3D Modelling of Human Hand with Motion Constraints

[1]Nishank Parashar, [2]Rakesh Soni, [3]Yash Manchanda, [4]Tanupriya Choudhury
[1]nishankparshar@gmail.com, [2]rakesh.soni1998@gmail.com, [3]yashm99999@gmail.com,
[4]tanupriya1986@gmail.com
[1,2,3,4] University of Petroleum & Energy Studies (UPES),Dept. of Informatics, School of Computer Science, Dehradun

Abstract: This paper describes the gestures and features of human hand motion, by considering the natural motion constraints of a hand. There are many restrictions that cannot be represented straight forwardly. Our aim is to reduce motion constraints by proposing some learning approaches to analyse and collect the dataset of motion behaviour of hand. The learning model involves application of Heat Map technology, along with usage of machine learning algorithms. We will implement some data mining algorithms to improve the accuracy of results and reduce redundancy in the testing data of hands space. This will result in better forecasting and prediction capability. Also we are using a microprocessor chip which will use sonar/radar for capturing ultrasonic waves, generated due to movement done by hands.

Keywords: Gestures, hand motion, Constraints, Dataset, Heat-Map, Ultrasonic waves, Data Mining, Sonar, CNN algorithm.

1. INTRODUCTION

In the recent years, several efforts have been devoted to recognize human hand gesture system and related motion analysis. In order to facilitate HCI (human computer interaction) and make computer more human friendly, we intend to make a system that could understand and decode human hand gesture by observing the minutest of joint motion in human hand.
Interactions provided by the traditional input devices such as mouse and keyboard are still behind the time, and are unable to provide natural interface through which we could gesture recognition system will prove to be a great advantage to mankind. It will also prove to be beneficial for person with disabilities. The deaf and mute could also use this technology without any restrictions.
The data glove technology will only restrict us to a specific kinematic model design by a manufacturer of hardware system, so this is not a recommended technology for capturing human joints motion.
Therefore to track the motion of human hand we need a better technology which could capture the motion precisely. We will be developing a high end image based on motion capture and analysis system. It will primarily include two main techniques -

Using Heat Map, and joint positions in the image also finding its probability distributions.

- By this we can generate the location of three dimensional joints by using the depth image
- A novel three dimensional regression strategy is proposed utilizing multi-view CNNs, so that depth cues can be exploited.
- Also, we are anticipating the profundity picture onto three symmetrical projection planes (x-y, y-z and z-x planes of the world arrange framework).

Detection of Ultrasonic waves, the principle of Doppler Effect is used here

- Ultrasonic is generally refers to anything above the frequencies of audible sound, and so includes anything over 20 KHz and can be extend up to 10 MHz and beyond.

978-1-5386-7709-4/18/$31.00 ©2018 IEEE

- Any movement of hand near ultrasound wave, will be received by sensor and will start predicting the gesture on the basis of data stored.

2. LITERATURE SURVEY

Not many works in research are able to address this issue of three dimensional hand picture approximation from a solitary two dimensional picture. This issue, somehow, limits our ability to interact with technology. Numerous methodologies have been proposed for hand gesture estimation, but still they are not fit for perfect use.

Model-dependent

Visual-based extraction of articulated hand is a tough problem due to several degrees of freedom.

In [1], a low dimensional model is created in which hand articulation were represented by set of linear manifolds, which leads to better approximations. But, this model is view dependent and valid only for the orthogonal view to the palm.

In [2], a comprehensive model is created by making a well versed dataset including several hand pose estimations. It consists of synchronized videos from various camera views. The shortcoming in this model is that the true hand motions still somehow remains undetected. Deriving the location of joints is not an easy job.

Search-dependent

Search dependent strategies follow non-parametric approach and involves finding closest neighbor search issue from vivid and large datasets, in which results are highly unstable because their dependency on low [3] or high [4] level highlights extracted from the picture.

A simple approach is depicted in [4] where two dimensional pose estimation is followed by three dimensional exemplary matching. It makes use of modular training in which two dimensional datasets are utilized further to train initial image processing.

Two dimensional Pose Estimation to three dimensional Pose Estimation:

Prior strategies toward this path take in pose conditioned joint angle limits and prepare a three dimensional models from dataset and recover the three dimensional depiction by reconstructing the two dimensional key-points [5]. The major drawback with these estimations is that it is not able to reduce the ambiguities in joint angle limits. During motion, it is possible that the hand reaches such a state that may not be possible to formulate statistically.

Three dimensional Pose Estimation to Images:

[6] Involves kinematic model fitting, which is more robust to obstructions, and includes three dimensional hand pose approximation from monocular RGB. This model cover several dimensions. It utilizes the secondary inputs like, depth images or multi-view RGB, which facilitates hand motion tracking in three dimensional. It is a better way to detect three dimensional hand pose estimation due to substantial synthetic data which almost resembles the real hand images.

A conditional generative model is obtained in [7] that learned approximate inference by training supplementary network. It utilizes GAN to change over artificially produced hand pictures to look more practical.

The major drawback with theses estimations is the overfilling problem. It requires a lot of training data so that it could predict accurately, which may become infeasible in some situations.

Some work related to micro hand gesture recognition has been done.

A project named uTrack used different types of wearable figuring turn out to be increasingly ordinary, the requirement for a decent component for ceaseless pointing will end up basic. [9].

N. Patel et al created WiSee, can empower wholehome motion acknowledgment utilizing few flag sources. Our outcomes in a 2-room condo demonstrate that WiSee can remove a rich arrangement of motion data from remote flags and empower entire home motion acknowledgment utilizing just two remote sources set in the lounge room. [10] These works excited more enthusiasm for using radar flag preparing in HGR. However the above work essentially centered with moderately wide-range and extensive move human signals.

In maximum radar signals, focal freq. & transmission limit the segment properties, they are not fit good for seeing little scale hand movements with inconspicuous developments in a few fingers, as they generally can't perceive fingers. The limitations in range and speed objectives are gigantic troubles for smaller scale hand motion acknowledgment.

Google Soli have proposed the principal motion acknowledgment innovation equipped for distinguishing a rich arrangement of dynamic signals in view of high-recurrence, short-extend radar. Our procedure is in view of a start to finish prepared mix of profound convolutional and repetitive neural systems. The calculation accomplishes high acknowledgment rates (average 87%) on a testing signal set counting eleven signals & crosswise over ten clients. [11].

3. TECHNICAL OVERVIEW

Radar is an abbreviation for Radio Detecting and Ranging. The name itself proposes that the radars are utilized to distinguish the nearness of item and decide its range, i.e., separation and bearing, utilizing ultrasonic waves.
The (RF) radio-frequency essentialness is passed to and get reflected back from the source. A little part of the waves which gets reflected and get added in the radar set. This returned essentialness is called an Echo, correspondingly, everything involved in sound phrasing. Radar sets choose the partition and heading by using echo of the reflecting thing. [12]

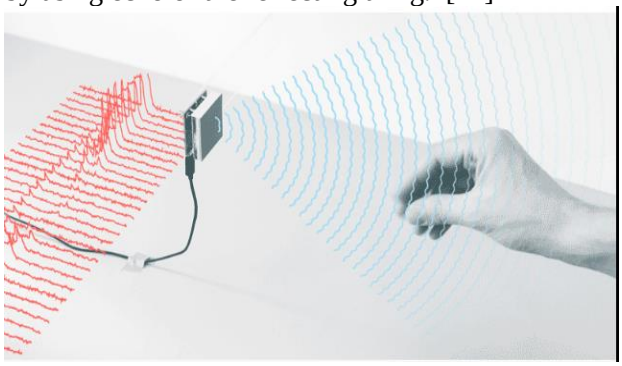


Figure 1: Radio Detecting and Ranging

Ultrasonic Waves and its classification:

A Support Vector Machine (SVM) is a discriminative classifier formally portrayed by an isolating hyper plane. All things considered, given named preparing information (supervised learning), the computation yields an ideal hyperplane which arranges new models.

To classify set of gestures we choose Support Vector Machine (SVM) classifier. SVM algorithm is demonstrated to accomplish a decent execution for real-world applications and with scientific models that depend on basic thoughts and are anything but difficult to dissect. Movement Frames got from back to back squares after some time shape a Motion Profile. On average a motion involves 2 seconds in time, producing a Motion Profile of 100 casings. Each motion type has its remarkable Motion Profile. [13]

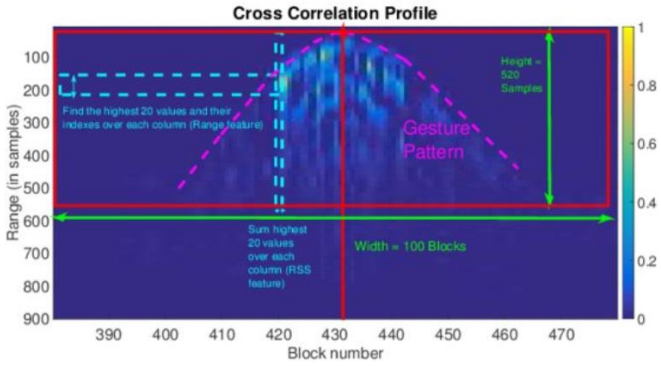


Figure 2: Ultrasonic Waves Classification

Heat Map and its Classification:

In neural systems, ConvNets or CNNs (Convolutional neural networks) is essential characterizations to do pictures acknowledgment, pictures groupings. Articles discoveries perceive faces that are a bit of the domain where Convolutional neural networks are commonly come in used.

CNN picture arrangements take an info picture, process it and characterize it under specific classes (E.g. Puppy, Cat, Tiger, Lion). PCs see an info

picture as a variety of pixels and it relies upon the picture goals.

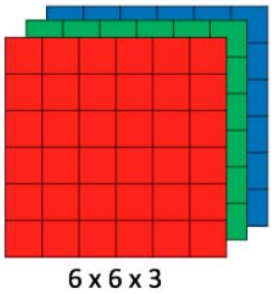


Figure 3: Array of RGB Matrix

In fact, profound learning CNN models to prepare and test, each information picture will go it through a progression of convolution layers with channels (Kernels), Pooling, completely associated layers (FC) and apply Softmax capacity to group an item with probabilistic qualities somewhere in the range of 0 and 1. The beneath figure is an entire stream of CNN to process an info picture and arranges the items dependent on qualities.

Heat map regression is now a standard approach for two dimensional pose estimation since it allows to accurately localize the key points in the image via per-pixel predictions. Creating volumetric heat maps for three dimensional pose estimation. [14]

We certainly learn depth maps and heat map appropriations with a novel CNN design. We utilize a current two dimensional present estimation demonstrate [24] to initially get the heat maps of hand key indicates and feed them another CNN that relapses an authoritative posture representation and the camera view point.

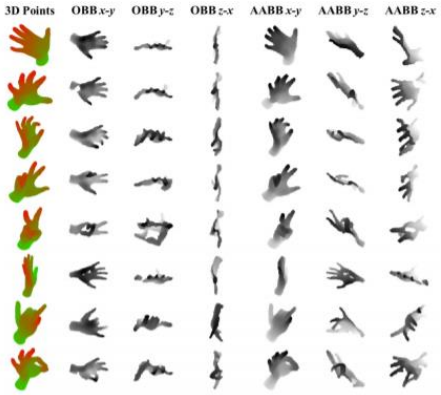


Figure 4: Heat Map Classification

4. RESULT ANALYSIS

The model which we have created is much more concise and is able to recognize the multitude of human hand gestures. Moreover, in order to achieve higher accuracy prediction results, we have executed the computationally more rigorous end-to-end network. As a result of this, we are able to get a classification accuracy of 96.25% using a more insightful Doppler Effect features. We have also implemented heat map regression to further increase our accuracy by another 1.32 %, achieving a near perfect result.

5. CONCLUSION AND FUTURE RESEARCH

In this paper, we proposed a novel approach to model the hand constraints. However, there is still much to be done to improve this model. Still, there are various hand motions which remain undetected. We can work on adding more sample to the database and make it more vivid for faster and accurate search results. We need to minimize the degree of freedom by finding out more constraints through deep study of human hand anatomy. Small radar sensors will enable the IoT by providing accurate intelligence to data aggregators. The conversion of two dimensional pose estimation to three dimensional pose estimation is a relatively time consuming process. Methods can be introduced to fasten this process of conversion. There is also a possibility of getting conflicting results as we are using two different classification algorithms for ultrasonic wave classification and heat map regression respectively. Research on

reinforcement learning and Artificial Intelligence and also using a single classification algorithm for image and wave classification can enhance the efficiency of this result. A mechanism for error detection for hand pose estimation could also be developed. By implementing this model, we can facilitate the human computer interaction for visually impaired and handicapped.

REFERENCES

[1]Wu, Y., Lin Et al.: Capturing natural hand articulation. In: ICCV. (2001)

[2]Sigal, L., Et al.: Synchronised video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. (2009)

[3]Romero, Et al.: Hands in action: real-time three dimensional reconstruction of hands in interaction with objects.

[4]Chen, C., Ramanan, D.: three dimensional human pose estimation = two dimensional pose estimation + matching. (CPVR)

[5]Akhter, I., Black, M.J.: Pose-conditioned joint angle limits for three dimensional human pose reconstruction
(ECCV)

[6]Mueller, Et al.: GANerated hands for real-time three dimensional hand tracking from monocular RGB. (2017)

[7]Goodfellow, Et al.: Generative adversarial nets.(2014)

[8]Umar Iqbal1,2 , Et al. Hand Pose Estimation via Latent 2.5D Heatmap Regression (ECCV),(2018)

[9] K.-Y. Chen, K. Lyons, Et al. ‘‘uTrack: three dimensional input using two magnetic sensors,’’ in *Proc. 26th Annu. ACM Symp. User Interface Softw. Technol.*, 2013,.

[10] Q.Pu,S.Gupta,S.Gollakota,andS.Patel,‘‘Whole-homegesturerecogni- tion using wireless signals,’’ in *Proc. 19th Annu. ICMCN* 2013, pp. 27–38.

[11] S. Wang, Et al., ‘‘Interacting with soli: Exploring fine-grained dynamic gesture recognition in the radio- frequency spectrum,’’ in *Proc. 29th Annu. Symp. User Interface Softw. Technol.*, 2016.

[12] Google Project Soli (2018)

[13] (PDF) Hand Gesture Recognition Using Ultrasonic Waves. Available from: https://www.researchgate.net/publication/320609351_Hand_Gesture_Recognition_Using_Ultrasonic_Waves [accessed Jan 08 2019].

[14] Hand Pose Estimation via Latent 2.5D Heat map Regression. Available from: https://arxiv.org/pdf/1804.09534.pdf [accessed Jan 08 2019].

[15] Wei, Et al., Convolutional pose machines. In: CVPR.

3D Printing: Factors Influencing its Quality and Nature

[1]Prakriti Saini,[2]Dhruv Garg, [3]Tanupriya Choudhury
[1]prakritisaini@gmail.com,[2]dhruv98garg@gmail.com,[3]tanupriya1986@gmail.com
[1]Indira Gandhi Delhi Technical University for Women, Kashmere Gate, New Delhi, Delhi
[2,]Amity University, Uttar Pradesh,Noida,India, [3] University of Petroleum & Energy Studies (UPES),Dept. of Informatics,School of Computer Science,Dehradun

Abstract- 3D(Three Dimensional) Printing is a kind of added substance fabricating innovation where a three dimensional question is made by setting down progressive layers of material which frames the last object. 3D Printers offer item architects the capacity to print parts and segments that are produced using distinctive materials which have different mechanical and physical properties in a solitary form process. In spite of the accomplishment of 3D printing there are sure irregularities which are looked amid the procedure, which will in general debase the nature of the print. In this report the point is to discover diverse components that influence the nature of a 3D print and its causes. Analyses were finished utilizing UTM to discover the impact on rigidity of the print under various conditions. For directing this assignment essential learning of Strength of Materials, Universal Testing Machine and 3D Printing is significant. Subsequently starting writing overview of the related subjects was done. The outcomes dependent on hypothetical reviews and exploratory outcomes are aggregated in this report.

KEYWORDS: 3D printing, mechanical, designs, UTM

I. INTRODUCTION

3D printing is a kind of added substance fabricating innovation in which a three dimensional question is made by setting down progressive layers of material which shapes the last object [3]. 3D printers offer item planners the capacity to print parts and segments that are produced using distinctive materials which have different mechanical and physical properties in a solitary form process. The further developed 3D printing innovations right now yield models that intently copy the appearance and usefulness of the last item. [4] 3D printing is accomplished by utilizing an added substance process, where progressive layers of material are set down in various shapes.[1][2][6]

Quality of Material:

Experience demonstrates that any material exposed to a heap may either twist, yield or break, contingent on the greatness of the heap, the nature of the material and its cross-sectional zone.

The mechanical properties of a material characterizes the conduct of materials under the activity of outside powers called loads. They are a proportion of the quality and enduring attributes of a material in administration, and are if an incredible significance in the plan of apparatuses, machines and structures. Mechanical properties are auxiliary delicate as in they rely on the precious stone structure and its holding powers, and particularly upon the nature and conduct of the blemishes which exist inside the gem itself or at the grain limits.

Universal Testing Machine**:** A UTM is a machine which is used to measure the stresses, generally tensile and compressive stresses developed in a specimen. UTM are capable of performing different types of tests on vast variety of structures. Most UTM can be adapted to fit the customer's needs.

II. EQUIPMENT UTILIZED

3D Printer: A 3D printer works on Additive Manufacturing which is often abbreviated as AM. AM is a process in which a three dimensional solid object is made by a digital file of the contour of the object to be printed. Progressive layers of melted material is poured by the 3D printer which shapes the object.

Fused Deposition Modelling (FDM): Fused Deposition Modelling was first found by Scott Crump who is also the founder of Stratasys. This technology works with specialized 3D printers and thermoplastics which are used for production ,to build parts or contours that are durable, strong and also dimensionally stable. A very high accuracy and repeatability is required in this process. The advantage of this technology is that it is simple to use and office friendly. The thermoplastics which are used in this technology are mechanically stable

and also environment friendly. It is easier to produce complex geometry with this technology.

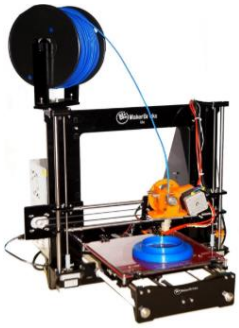


Fig1. Makerbricks I3C 3D Printer

Makerbricks I3C 3D Printer

The printer used for the present work was I3C 3D Printer by Makerbricks. It is a very simple 3D printer which is based on the FDM technology. The maximum build up volume is 180mm X 180mm X 180 mm. The printer uses a nozzle diameter of 0.5mm, but the nozzle is easily replacable.

Ponterface software was used to provide required instructions to the printer from the computer. The 3D model was generated in Solidworks 2012 and the .stl file was exported to slicing software, Cura 2.3.0.

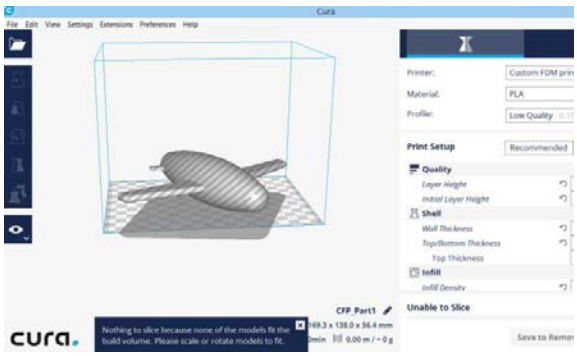


Fig2. Snapshot of cura2.3.0 GUI

III. MATERIAL USED

PLA is one of the most suitable material used in 3D printing. The filament of PLA is heated and then the successive layers of specimen are printed. Every specimen has its unique characteristics.

IV. QUALITY OF 3D PRINTING

Geometry:

Geometry of a 3D printed object can reduce its quality. Although it is said that 3D printing is shape complexity free but the quality of a 3D printed product also depends on the geometry of its shape. Geometry that requires support often result in poor surface quality of the product.

Material Properties:

Material's properties such as its melting point, texture, cooling point, fusion rate etc. are also taken into account for a good quality product. Generally, the materials that are used in the industry are PLA (polylactic acid), ABS (acrylonitrile butadiene styrene) and PVA .

PLA is used because it is a biodegradable material which is extruded at a low temperature and does not require a heated bed. PLA is stiffer than ABS. It is a brittle material used for manufacturing cosmetic prints, prototypes, desk toys etc.

ABS is the cheapest of the three. It is best extruded at a temperature range of 215º-250ºC and requires heated bed which helps in preventing warping. This ABS plastic material is less brittle and more ductile in nature.

It is important to note that all materials are not compatible with all printers, as the temperature range of extrusion ranges from 160º-305ºC, a printer may really get worked up due to high temperature of extrusion. [4]For example, if a printer is designed for PLA, it is best extruded at a temperature up to 250ºC, may totally fail at 300ºC. Hence, the property of a material is very important while computing quality of a product. Also properties like mechanical strength and fatigue life must be taken into consideration while choosing a material.

Temperature of cooling settings:

As we all know that plastic shrinks as it is allowed to cool down at a lower temperature. While printing the first layer, in many materials like ABS (acrylonitrile butadiene styrene), needs printing bed to be maintained at a particular temperature if not done so it will lead to shrinkage of the plastic material. Due to the temperature difference between the two materials, the print bed and the plastic, the plastic will tend to separate from the print bed as it cools.

High Temperature:

If the temperature of the extruder is too high the plastic inside the nozzle will absorb more heat and become less viscous. As the filament in the nozzle loses its viscosity it becomes easier for it to flow out of the nozzle, it can often lead to leaking out of the filament. This leakage of the filament due to high extrusion temperature generally results in

defects like stringing of think filament threads on the product printed.

Warpage due to overheating is another defect which occurs due to extrusion of filament at very high temperatures.

Low Temperature:
As mentioned in 4, If the extruder temperature is too high the filament tends to lose its viscosity. If the extrusion temperature is too low the filament might not have melted properly and still be in the solidus range. This means that the filament contains solid as well as liquid particles. Either these solid particles will clog the nozzle and no extrusion will be possible then or the extrusion will not be proper due to insufficient filament extruded. This will degrade the quality of the printed product. One of the defects which are seen in the product due to low extrusion temperature is stringing or oozing of the filament.

Another defect caused due to low extrusion temperature is layer separation and splitting. It is a well known fact that warm plastic will always bond together better than the cold plastic. Therefore, if the extrusion temperature is too low the two alternate layers of plastic will not show proper fusion and will result in poor mechanical strength[5].

Rate of cooling:
Hot plastic is less viscous. It has tendency to flow easily as compared to cold plastic. Therefore, rate of cooling of plastic must be maintained to ensure the proper solidification of the extruded plastic. If rate of cooling is not accordingly maintained the plastic will flow regardless of the shape required of the product and result in distorted figure of the desired product.

It is observed that often increased cooling rates result in formation better overhangs and bridges.

Size of the filament:
Using the correct diameter size of the filament is very important in 3D printing. The intake of filament of a 3D printer is measured in length of the filament and not volume. Hence, one of the most common errors seen is using wrong size of the filament. [6]If the diameter of the filament that you are using is very small as compared to the size of the nozzle then the plastic extruded from the nozzle will not be sufficient for the product to be printed. This will result in gaps in between the product.

Softwares which are used for 3D printing does not tell you how much plastic has actually left the nozzle. It might be possible that the plastic exiting the nozzle is lesser than what is calculated by the printer software. If this happens, you may start to notice gaps between adjacent extrusions of each layer. Therefore, it is very important to check whether your software knows the correct value of the filament diameter.

Build Platform Surface:
Build platform is the base on which the first layer is to be printed. It is very important to have a proper build platform surface for the first layer to stick on the bed properly. Sometimes the build platforms are not leveled. This can be a major fallout in 3D printing. It is observed that if the Build Platform Surface is not leveled properly the print does not stick to the layer bed.

We know that different plastic tends to adhere better to different materials, for example; PLA adheres better to BuildTak sheet and ABS adheres better to heat treated glass, such as Borosilicate. So, the material used for print bed also plays a major role in determining the quality of the product.

Low Infill percentage:
Infill percentage means the percentage of the product that is solid, i.e. the remaining percentage of the product is hollow. [7]Note that the infill on the inside of the product will be acting as the foundation for the layers above it. The solid layers which you will be printing will be printed on the top of these layers, using them as their foundation. If our infill percentage is too low there will be large gaps between your infill.
For example, if you have 20% infill, this means that you have to print the solid layers at the top on the foundation which is 80% hollow.
Notice that, the more infill will lead to cleaner print as well as better mechanical properties, whereas less infill will lead to poor quality of the product.

Layer Height:
It is a general rule of thumb that the layer height that you choose for the print must be 20% smaller than the diameter of the nozzle, this makes sure that the new layer of plastic which is being extruded is pressed against the layer below so that they can bond together properly. Most printers have diameter between 0.3-0.5mm. Taking an average of 0.4mm diameter nozzle, the layer height can be of maximum 0.32mm. If the layer height goes past

this, the adjacent plastic layers will not stick properly and can split or get separated easily.
Small layer heights are better for overhangs and bridges.

Use of primer:
When idle at a very high temperature, extruders start leaking plastic. The hot plastic oozes its way out of the nozzle tip. This then creates a void inside the nozzle from where the plastic is removed. Now, when you start printing again the printer will take some time before it starts extruding the plastic again. To prevent this, the extruder is primed properly so that the printer can start extruding as soon as it starts.

Extruding too much plastic:
As stated above, the printer has no software to evaluate the value of actual plastic that has been extruded. Sometimes when the temperature of the nozzle is too high the plastic oozes out of the nozzle, that is when too much plastic is extruded. When this happens each layer is slightly thicker than intended. Too much extrusion of plastic from the nozzle can be due to very high nozzle temperature or due to improper software settings.

Extra plastic that is extruded from the nozzle leads to too many scars on the top layer of the product. This reduces the quality of the product.

Dust Particles:
There are millions of dirt particles present around us, in the air we breathe. But what effect do they have on the quality if 3D print?
Dust particles which are present on the filament gets accumulated inside the nozzle during intake of filament. These particles which are accumulated inside the nozzle tends to block the nozzle. This will result in under or no extrusion.
Note that the plastic is heated up to its melting temperature inside the nozzle, the dust particles which are accumulated inside the nozzle gets mixed up with the melted plastic resulting in impure plastic, which further gets extruded according to the print required. This mixing of impurities in the plastic degrades the quality of plastic and its mechanical properties.

Nozzle too close to the print bed:
It is necessary to maintain a certain distance between the print bed and the nozzle to let the plastic extrude fluently. If the nozzle is placed too close to the print bed due to lack of space it will be difficult for the plastic to extrude properly. The hole at the end of the nozzle will be blocked and no plastic can escape. To prevent this make sure that there is some space between the print bed and the nozzle before starting to print.

Nozzle too far from the print bed:
At times the distance between the print bed and the nozzle is very huge than required. This large distance between the nozzle end and the print bed is not desired as it affects the quality of our product. Ideally the filament used must be slightly squished against the build plate to ensure proper adhesion if initial layers on the print bed. If the nozzle it too far from the print bed and not squished against it, there is a possibility that the plastic might not adhere effectively to the build plate.

Size of the Nozzle:
The diameter of the nozzle used in a printer also plays a major role in deciding the quality of the product. It is very clear by our previous discussions that for a good surface finish it is necessary to have a small layer height. This can be achieved by having small size of the nozzle, i.e. the smaller the diameter, better the finish. Hence, the printers that uses small diameter of the nozzle result in better surface finish. Small sized diameters will result in more detailed print. One of the major drawbacks of using a small sized nozzle is that it takes longer to print. Small nozzle size will decrease the rate of production of a product which is a major flaw in industrial sector.

Retraction Distance:
Retraction is the process where the filament is recoiled back into the nozzle to prevent it from dripping at the areas where no plastic is required. Retraction distance is a setting in slicer software which determines the amount of plastic that is pulled out of the nozzle.

If the amount of plastic retracted by the nozzle in more the less likely nozzle is to ooze out plastic. This will prevent stringing of the product. Retraction is not required in all materials. In some materials like TPE retraction process is deactivated to ensure that the filament has not coiled on the extruder pinion.

Retraction speed:
Retraction speed is how fast the filament can be retracted from the nozzle. If you retract too fast, the filament present may separate from the hot plastic that is present inside of the nozzle. This will result in breaking of the filament wire.

If you retract too slow, the hot plastic inside the nozzle will slowly start to ooze down. After a while it may start leaking through it even before the extruder has moved to its new destination. This is another cause of stringing of plastic on the printed products.

Printing at a very fast speed:
We know that by increasing the print speed of the machine will reduce the time consumed for printing of a single product. This is also helpful in preventing the oozing of plastic from the nozzle.

On the other hand printing at a very fast speed has its disadvantages as well. Due to high speed of the machine the plastic does not get enough time to extrude from the nozzle properly leading to gaps between the infill and outline. Ringing is another defect which is caused due to high printing speed. Ringing is a wavy pattern that appears on the surface of the print due to wobbling of the printer. Generally, wobbling of the printer is caused by operating the machine at a very high speed.

If you are printing too quickly there is a possibility that it might not allow the first layer to cool properly. This will result in improper fusion of the consecutive layers and we will get a distorted or warped product.

Support Infill Percentage:
There are geometries that consist of overhangs and bends. If you have a steep overhang or bend and nothing below it, supports can provide foundations to these layers. It is better to build a support for such shapes to prevent them from distortion and get a desired final product. Supports are built simultaneously by the printer. Slicer software provides such benefits to create innovative support structures to create complex shapes.

Slicer software allows you to decide the infill of the support structures by changing the Support Infill Percentage. Generally support infill is kept between 20-40%. If the bottom layers of your part are drooping too much it is better to increase the support infill percentage. Alternatively, you can also use lower density for the majority of the support structure and high infill percentage near the top of the supports. This will reduce the consumption of support material.

Residual Stresses:
Rapid heating and cooling or expansion and contraction is a cause of residual stresses that arise in material used for 3d printing process. In any case where the residual stresses exceeds the value of tensile strength of the specimen, a defect like cracking or warpage of the substance occurs. Hence, it is important to prevent the development of residual stresses inside the structure.

V. TESTING USING THE PROCEDURE OF TENSILE STRENGTH

- Note down the initial gauge length of the specimen. The gauge length should be symmetrical to the length of the bar.
- Complete the upper and lower chuck assemblies by choosing appropriate jaw inserts. Also, to make sure the smooth motion, apply some graphite grease on the tapered surface of the tip. Handle the upper end of the test piece tightly in the jaws of upper cross head. Similarly, lift up the lower cross head and firmly hold the lower part of the test piece in the jaws. When the specimen has been gripped vertically adjust the UTM such that it reads zero.
- By turning the capacity change wheel which has ram at the bottom of its stroke one can select the chart range required. By the quick setting control raise the ram by a few mm and set zero.
- Attach the extensometer tightly to the specimen and then adjust it so that it reads zero.
- Switch ON the UTM.
- UTM increases the load gradually until the yield point is reached, read the extensometer and load recordings and note them down carefully.

VI. TENSILE TEST OBSERVATIONS

Specimen Table-1

Length: 190mm Width: 19mm
Height:10mm

Load (N)	Extensometer (mm)	Stress (N/m^2)	Strain
0	0	0	0
30	4.3	157894.7368	0.022632
36	5.6	189473.6842	0.029474

Specimen Table-2

Length:190mm Width:19mm
Height:10mm

Load(N)	Extensometer(mm)	Stress(N/m^2)	Strain
0	0	0	0
18	2.4	94736.8421	0.012632
36	4.7	189476.864	0.024737

Specimen Table-3

Length:170mm Width:19mm
Height:7mm

Load(N)	Extensometer(mm)	Stress(N/m^2)	Strain
0	0	0	0
26	4.7	195488.721	0.027647
42	6.8	315789.473	0.04

Specimen Table-4

Length:170mm Width:19mm
Height:7mm

Load(N)	Extensometer(mm)	Stress(N/m^2)	Strain
0	0	0	0
11	3.4	82706.7669	0.02
22	4.5	165413.533	0.026471

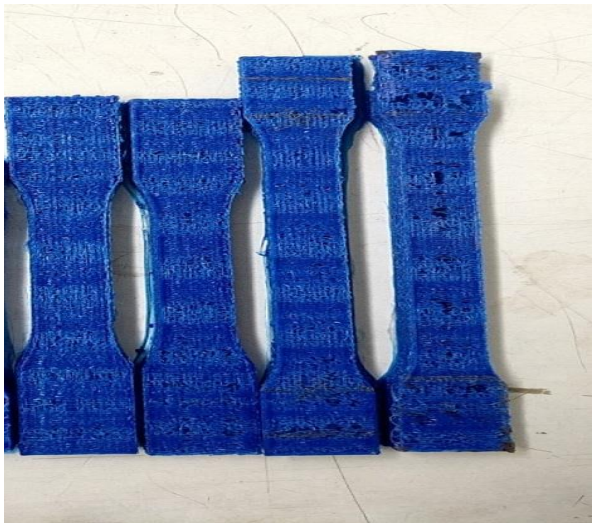


Fig 3. Specimen before Tensile Strength Test

Fig 4. Specimen after Tensile Strength Test

Specimen Specification Table

Serial No.	Layer Height (mm)	Top And bottom layer (mm)	Shell Thickness (mm)	Print Speed mm/s^2	Print Temp °C	Bed Temp °C	Infill Density
Specimen1	0.2	1	1	60	200	60	20%
Specimen2	0.2	1	1	60	200	60	50%
Specimen3	0.2	1	1	60	200	60	40%
Specimen4	0.2	1	1	60	200	60	60%

VII. RESULT, CONCLUSION AND FURTHER SCOPE

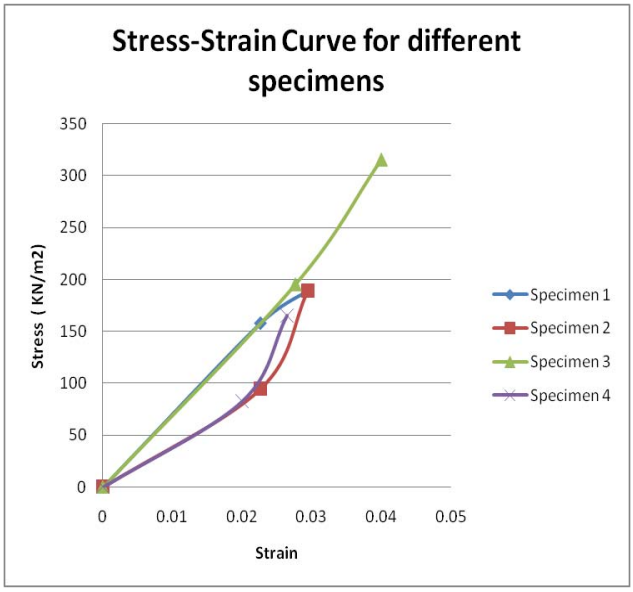


Fig 5. Stress-Strain graph of the four specimen

3D Printers are considered remarkable because they can produce complex geometry and shapes of an object by the same material, using the same machine. 3D printing technology is where a three dimensional object is created by setting down the progressive layers of melted material that builds up to shape the desired object.

The most common defects seen in 3D print are also listed in this project. Factors that affect the quality of a 3D print are stated above. In this project, by keeping all the parameters same and varying the infill density of the specimen it was observed that the Ultimate Tensile Strength of specimen 1 and specimen 2 are approximately equal. This shows that an object with 20% infill density will show nearly the same strength as an object with 50% infill density. Hence, by using 20% infill in formation of an object the strength and the quality of the object will remain the same and the consumption of material will be reduced.

References

[1] Simply3D.com,2017

[2] Samer Mukhaimar, Saed Makhool, "3D Printing Technology", Faculty of Engineering and Technology,December 2014

[3]Christopher Barnatt ,"3D Printing-The next industrial revolution",2013

[4] Raghav Bansal, Gaurav Raj, Tanupriya Choudhury, Blur image detection using Laplacian operator and Open-CV,SMART,2016.

[5] Hamzah, HairulHisham; Shafiee, SaifulArifin; Abdalla, Aya; Patel, Bhavik Anil ,"3D printable conductive materials for the fabrication of electrochemical sensors: A mini review",Electrochemistry Communications,2018

[6] Taufik, Mohammad; Jain, Prashant K., "Role of build orientation in layered manufacturing: a review",International Journal of Manufacturing Technology and Management, 12 January 2014

[7]Bin Hamzah, HairulHisham; Keattch, Oliver; Covill, Derek; Patel, Bhavik Anil,"The effects of printing orientation on the electrochemical behaviour of 3D printed acrylonitrile butadiene styrene (ABS)/carbon black electrodes",2018

A Computation Classified Analysis Based on Socio Economic Cost for University Students Accommodation

[1] Sugandha Gupta, [2] Tushar Pandey, [3] Tanupriya Choudhury, [4]Amrendra Tripathi
[1,2,3,4]University of Petroleum and Energy Studies,Dehradun,India
[1] anshusugandha@gmail.com, [2] tusugandha26@gmail.com, [3] tanupriya1986@gmail.com, [4]tripathiamrendra@gmail.com

Abstract – The research aims to analyse the accommodation facilities that the students of Delhi University use in absence of proper hostel facilities. It tries to find the relation between the choice of accommodation and factors like income and gender. Further, so as to make an estimate of the social cost involved we have tried to find out if the outstation students face any discrimination on account of their language, caste, creed or gender. The data has been collected with the help of questionnaires and the responses have been analysed quantitatively using tools like, bar graphs, pie charts and column charts.

Keywords: Analysis, Evolution, Distribution, Accomodation

I. INTRODUCTION

EVOLUTION OF STUDENT HOUSING INDUSTRY IN INDIA

With growth in the real estate sector in India new asset classes are emerging, in addition to residential, commercial and hospitality asset class. One such alternative that has developed over the past few years is, Student Housing. With knowledge being identified as the driver of the Indian Economy, investment is constantly being made in the education sector by both government and private institutes and as the people are understanding the importance of education, an increase in the number of students going out of their native place, is being observed. Earlier, these students were provided accommodation in the college premises itself, but now, due to an increase in number of such students, especially in private universities and colleges, it is becoming impossible to provide a place to stay to each and every student. Due to an increase in demand in this sector by students, the student housing industry is constantly flourishing in India.

UPES AND OUTSTATION STUDENTS

A "University" is defined as a higher-level educational institution in which students study for degree and where academic research is done. UPES one of the premier institutes of India imparting higher education to the students in different disciplines. Every year thousands of students come to UPES in hopes of getting admission in one of its colleges and it is indeed a tough job to make it through the high cut-offs. Those who however manage to make it through are faced with the problem of accommodation, which proves to dig a bigger hole in the pockets of the students than the expenditure on their studies does. With a constant increase in applicants from outside dehradun, it is becoming impossible for the university to provide accommodation facility to all the students. Constantly, the university has been considering construction of more hostels so as to relieve the students of the problem of lack of facilities of accommodation provided by its affiliated colleges.

ACCOMODATION FACILITIES AVAILABLE FOR UPES STUDENTS

Of the two branches of colleges law college and school of science, the college offers only 2 boys and 2 girls on-campus hostel facilities for undergraduate students. Besides, there are other 19 hostels for both undergraduate and postgraduate students. These hostels can accommodate a maximum of around 9,000 students whereas 56,000 students take admission to the university every year.

The off-campus hostels include around 20 girl's boys hostel and around 50 boys hostel with minimum 50 students in each.

Colleges that provide hostel facilities include Daulat Ram College, Hindu College, Kalavati Gupta Hostel, Indraprastha College for Women, Kirori Mal College, Lady Shri Ram College for Women, Miranda House, Hansraj College, Shri Ram College of Commerce and Sri Venkateswara College, DIT university, Graphic era university, UIT university. These hostels often provide seats to students on the basis of their marks and of the seats available, several are reserved under different categories and thus only a few students manage to get a university or college hostel. Moreover, of the few colleges that do have hostel facility, several have it either only for boys or for girls alone. In such a scenario, it is often flats or PGs that students rely on, with some of them providing fooding, while others do not even provide that facility.

II. REVIEW OF LITERATURE:

Grayson (1994) in his research showed that the place of stay severely impacted the performance of a student during the first years of college. According to him, the students who lived in campus performed better than the ones who live in off-campus facilities. This conclusion was drawn on the basis of primary data that he collected from across different universities regarding the overall annual performance of the students who live in and off campus. Garg et al. (2015) argued that , girls usually prefer private hostels and PGs while boys prefer to live in shared flats. They had conducted their study, comparing the difference in students housing market in four different cities in India, namely, Pune, Chennai and Delhi. This study was again done using primary data collected from major private and government colleges at these places that are known as the hubs of knowledge in India. Gupta et al. (2014) after conducting a study on the housing facilities in Bangalore city of India found that out of total number of outstation students, only 15% receive housing in college hostels. Owens (2010) proved through an empirical study that staying away from parents for higher studies in colleges have great impact on the psychology and personality of students. Girija (2015) conducted a research on the students of University of Delhi and University of petroleum and petroleum studies and using a random utility framework, estimated that women are willing to choose a college in the bottom half of the quality distribution over a college in the top quintile in order to travel by a route that is perceived to be one standard deviation safer. Furthermore, women are willing to spend INR 18,800 (USD 290) per year more than men for a route that is one SD safer an amount equal to double the average annual college tuition.

RESEARCH GAPS

A review of the existing literature makes it clear that there is a lack of proper study in various areas which are various factors affecting choice of accommodation by students, Analysis of the monetary cost borne by the students for accommodation, Social costs in terms of kinds of discrimination that outstation students face.

III.RESEARCH METHODOLOGY

The study has been carried out to analyse the various factors that affect the choice of accommodation by students of UPES and University of Delhi. As the study population consists mainly of literate students pursuing higher studies, a questionnaire was used to collect data. Thus, the study is mainly based on the information got from primary sources. The algorithm used in this simplification and calculations is decision tree algorithm. It belongs to the family of supervised learning algorithms. The general motive of using decision tree is to create a training model which can use to predict class or value of target variables by learning decision rules inferred from prior data(training data).[1][2][4]

DATA SOURCE

For the objective of analysing the relation between the income level of the family a student hail from and the choice of accommodation questions regarding the same has been collected through questionnaire. The specific questions that pertain to this particular study are: Income level of parents (Combined income of both the parents if both are working)

Place of stay- Hostel, PG, Flat, Others

For the objective of analysing how gender affects the choice of accommodation again, a questionnaire was used. The questions that were specific to the study of this objective are:

Gender- Male, Female, Others

Place of stay- Hostel, PG, Flat, Others

Most important factor affecting your choice- Cost, Proximity to college, Security, Food quality. Besides, the apparent monetary costs that the students explicitly, there are several social costs that go unnoticed. To analyse them, we tried to find about various discriminations that students belonging to different regions, religions, caste, creed and gender face. Again, a questionnaire was used to collect information and the questions specific to the study of this includes: Caste- General, OBC, ST/SC. Mention the religion if you are comfortable. From where do you belong? Gender – Male, Female, Others Deadline to reach place of stay, if any. Any kind of discrimination that you feel or discomfort that you feel you have to face because of your social background. Along with various implicit social costs that outstation students bear, there is a huge monetary cost involved in living in a different city, which dig a big hole in the pockets of such families that send their kids outside. To analyze this cost and the variation in the amount borne by students residing in hostels, PGs and Flats, the following questions had been asked: Rent paid per month, Expenditure incurred on food, Expenditure incurred on electricity, AC/Non AC Room, Are you satisfied with the food, Any travelling expenditure you have to bear on a daily basis to reach to college? If yes, how much? As most of the respondents living in PGs and hostels did not provide a separate information about rent, fooding and electricity, a separate analysis under this head could not be made. To compare the cost incurred mean cost incurred by respondents belonging to each group has been calculated. This mean cost has again been calculated on an annual basis just to make it easy to compare. Mean refers to the average that is used to derive the central tendency of data in the question. It is determined by adding all the data points in the population and then dividing the total by the number of observations.

SAMPLING AND CLASSIFICATION

The method of sampling used is random sampling from a population that consisted of outstation students studying in UPES. Google Forms were send by using social media platforms with clear instructions that the form is for the outstation students studying in Delhi University and UPES. A total of 70 responses were collected.

The responses received were further classified (using cluster algorithm to cluster similar students together)under various heads so as to achieve the objectives. The classification heads and the percentage of sample size lying under each head shall become clear from the following chart and table.

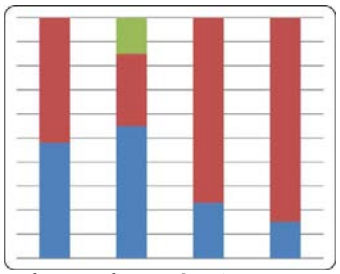


FIGURE 1: Classification of respondent (source: primary data)

TABLE 1: Classification of respondent (source: primary data)

BROAD CATEGORY	SUB CATEGORIES	No OF RESPONDENTS	PERCENTAGE
GENDER	MALE	72	48%
	FEMALE	78	52%
CASTE	GENERAL	83	55.3%
	OBC	45	30%
	ST/SC	22	14.7%
REGION	SOUTH INDIA	38	25%
	NORTH-EAST INDIA	6	4%
	OTHERS	106	71%
RELIGION	MUSLIMS	53	35.3%
	NON-MUSLIMS	97	64.7%

DATA ANALYSIS

To find the relation between income level and choice of accommodation the respondents were classified into four categories on the basis of their income level. The four categories into which they were classified include-

1. Income upto 2 lakhs

2. Income between 2 to 5 lakhs

3. Income between 5 to 10 lakhs

4. Income above 10 lakhs

Thereafter, the number of people living in PGs, Hostels and Flats under each category was recorded and a frequency distribution was made. Stacked column chart was used for proper presentation so as to easily analyse if there is any relation between income level and the choice of accommodation. Percentage of people choosing the alternative options under each category was found out so as to make it easier to compare. Again, the respondents were classified as per their gender and a frequency distribution was prepared to find out the number of respondents under each category choosing- PGs, Flats and Hostels. A clustered column chart was used to present the percentage of respondents in each category, lying under different choice heads. Percentage is used instead of absolute numbers so as to make a comparative analysis since the number of respondents lying in each category differs in size. To analyse the social cost if any involved, the respondents were asked if they feel that they were being subjected to any kind of discrimination [10][11]. The respondents answering in affirmative were asked to mention the kind of discrimination they faced. Since the various kind mentioned by the respondents included discrimination on the basis of religion, region and gender. An observation was made so as to see, from which place and religion do the people responding in affirmative belong. The findings have been summarised using table and pie charts. Since no respondents mentioned about the kind of discrimination being faced on basis of gender, an observation was made from the responses of people belonging to male and female category and the difference in responses observed to the question of deadline was summarised. Again, using a table and a pie chart. To analyse the difference in the cost incurred by residents of PGs, Flats and hostels the average annual cost of respondents belonging to each category was found. Since, some responses did not give a separate information about the cost incurred on food, electricity and rent, the aggregate annual spending of each respondent was calculated and the mean of these aggregates was found under each category. Annual spending was used for this category instead of domestic spending because most hoteliers mentioned the costs on annual basis. Mean refers to a measure of central tendency calculated by dividing sum of all observations by the number of observations. To properly present and analyse the data, tables and column charts are used. [3][5][6]

IV. RESULTS:

To find relation between income and choice of accommodation.

The respondents were asked questions about their income level and their respective choice of accommodation was recorded.

TABLE 2: Distribution of respondents on basis of income level

INCOME BRACKET	**NO. OF RESPONDENTS**	**PAYING GUEST**	**HOSTELS**	**FLATS**
Upto 2,00,000	15	3	2	10
BETWEEN 2,00,000-5,00,000	34	18	2	14
BETWEEN 5,00,000-10,00,000	49	27	5	17
ABOVE 10,00,000	52	40	5	7
TOTAL	150	88	14	48

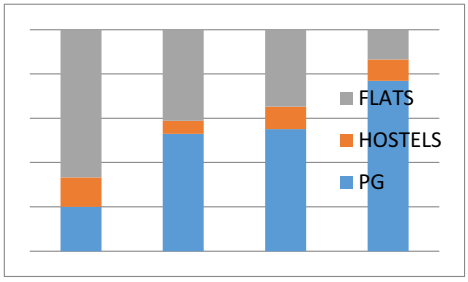


FIGURE 2: Distribution of respondents on basis of income level

According to the responses 58.66% students live in PGs, 9.33% live in hostels and 32.10% live in flats. Of the people belonging to the income group below 2lakhs, 20% live in PGs, 13.33% live in hostels and 66.67% live in flats. Of the respondents lying in the income group between 2 to 5lakhs, 53% live in PGs, 5.88% in hostels and 41.12% in flats. Of the respondents belonging to the band between 5 to 10lakhs 55% live in PGs, 10.20% in hostels and 34.80% in flats.[12]

To find if gender effects choice of accommodation

To find if gender effects choice of accommodation

TABLE 3: Distribution of respondents on basis of gender

GENDER	**PG**	**HOSTEL**	**FLATS**
MALE	30%	5%	65%
FEMALE	54%	7%	39%

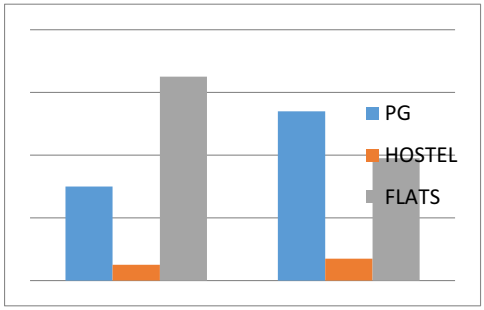


FIGURE 3: : Distribution of respondents on basis of gender

According to the responses, of the males, 30% live in PGs, 5% live in hostels and 65% live in flats. Of females, 54% live in PGs, 40% in flats and 6% in hostels. We thus conclude that there is indeed a relation between gender and choice of accommodation. While more girls prefer PGs, boys find flats to be a better option. The analysis of the question of factor playing most important role in the choice, as given below makes the reason for this clear.

TABLE 4: Distribution of responses on the basis of factor playing most important role in choice

GENDER	PROXIMITY TO COLLEGE	SECURITY	COST	FOOD QUALITY
MALE	9%	26%	43%	22%
FEMALE	13%	55%	12%	10%

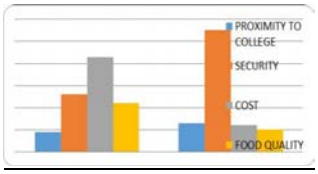


FIGURE 4: Distribution of responses on the basis of factor playing most important role in choice

According to the responses, 9% of males lay most importance on proximity to college, 26% lay it on security, 43% on cost and 22% on food quality. Of females, 13% lay most emphasis on proximity to college, a major 55% on security, 12% on cost and 10% on food quality. We see that majority of girls mark security as the most important factor that they consider while choosing their place of stay. It is quite clear that they find Paying Guests to be more secure than flats. Hostels again, being allotted on basis of merit, gets out of the equation.[7][8]

To find about social cost, if any involved

Students were asked if they face any kind of discrimination and their response to the same was recorded. The following tables and figures summarise the results.

TABLE 5: Classification on the basis of response to the question of discrimination.

RESPONSE		PERCENTAGE OF RESPONSES
NO		67% (101 of 150)
YES % OF THE AFFIRMATIVE RESPONSES		33% (49 of 150)
REGION	42% (21 of 49)	
RELIGION	27% (13 of 49)	
GENDER	31% (15 of 49)	
OTHERS	0% (0 of 49)	

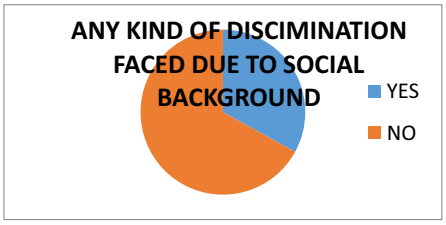


FIGURE 5: Classification on the basis of response to the question of discrimination.

Of the responses got 33% agreed to having faced discrimination of some kind, while 67% denied having faced any discrimination.

The above analysis clearly shows that though not much, there is a social cost that does exist for students living outside their home town. It is present in form of various kinds of discriminations that they face.

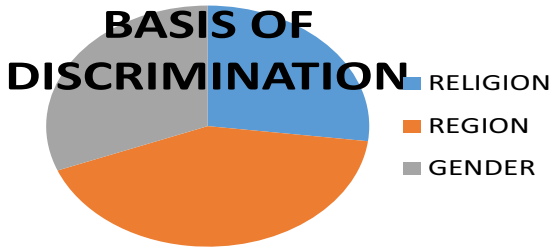


FIGURE 6: Classification of the affirmative response.

Of the people responding yes, 27% complained of discrimination faced on the basis of religion, 42% on the basis of region and 31% on basis of gender.

TABLE 6: Classification of responses on basis of region and religion

CATEGORY	NO. OF RESPONSES	PERCENTAGE
REGION NORTH-EAST INDIA SOUTH INDIA OTHERS	 6 15 0	 29% 71% 0%
RELIGION MUSLIMS CHOOSE NOT TO SAY CHRISTIANS	 7 5 1	 54% 38% 8%

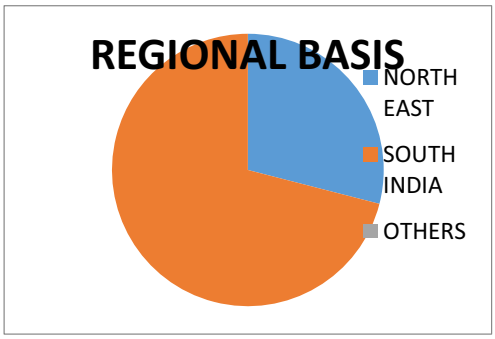


FIGURE 7: Distribution on the basis of region

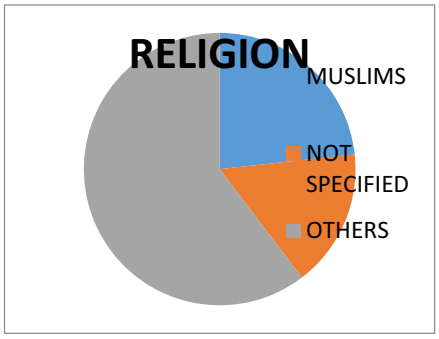


FIGURE 8: Distribution on the basis of religion

Of the people who responded in affirmative to the question of discrimination on basis of religion, 54% were Muslims, 38% chose not to reveal their religion and 8% were Christians. Out of the 13 responses received in this regard, only 6 specified the kind of discrimination, all of them referring to difficulty in finding place of accommodation due to their religion. Of the people who responded in affirmative to the question of discrimination on basis of region, 29% were from North-East India and 71% from South-India. Of the 15 responses received from the South Indians 9 specified the kind of discrimination as one being done on basis of language.[9]

TABLE 7: Distribution of responses on basis of gender, to the question of deadline

	YES	NO
MALE	17%	83%
FEMALE	72%	28%

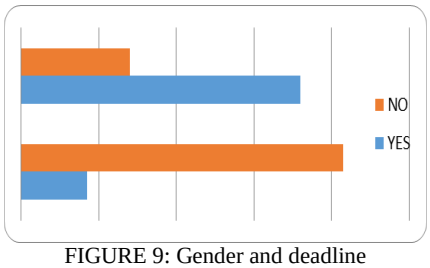


FIGURE 9: Gender and deadline

On the question of deadlines, while 83% male respondents replied in negative, only 28% of female respondents had no deadline being imposed on them. It is does observed that in the name of security of the girls, a deadline is being imposed on them as opposed to boys which is clearly against the right to equality. To analyse the overall monetary cost incurred on accommodation by students. Average cost incurred by respondents belonging to each category.

TABLE 8: Table comparing average annual cost incurred by students living in pgs, flats and hostels

AVERAGE ANNUAL COST	PG	FLAT	HOSTEL
(INCLUDING RENT, FOOD AND ELECTRICITY)	1,70,000	96,000	70,000

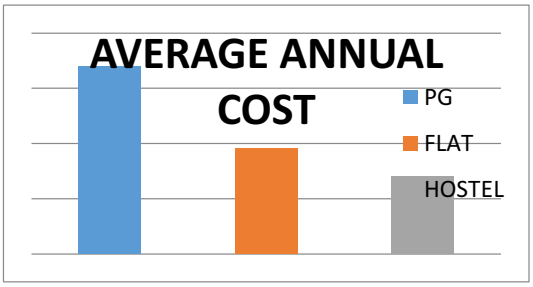


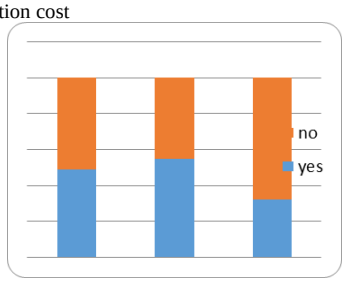
Figure 10: Table Comparing average annual cost incurred by students living in pg, flats and hostels

The responses show that students living in PGs pay the highest annual cost of around 1,70,000. The students living in flats and hostels on the other hand pay around 96,000 and 70,000 respectively. Comparison on the basis of facilities provided and advantages of choosing the alternative options

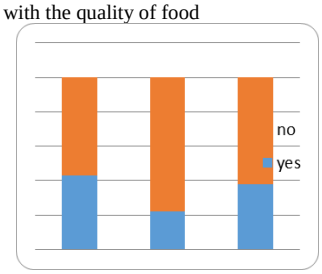
TABLE 9: Comparison on basis of facilities

	PGs	Flats	Hostels
Do you have to incur any transportation cost on daily basis to reach to your college? YES NO	 49% 51%	 55% 45%	 32% 68%
Are you satisfied with the quality of food you get? YES NO	 43% 57%	 22% 78%	 38% 62%
Do you have an AC Room? YES NO	 100% 0%	 19% 81%	 0% 100%
Do you have WIFI facility? YES NO	 100% 0%	 0% 100%	 92% 8%
Do you feel that WIFI facility is important? YES NO	 15% 85%	 22% 78%	 17% 83%

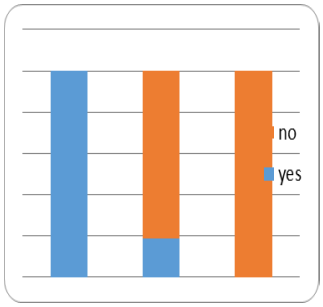
A).Transportation cost



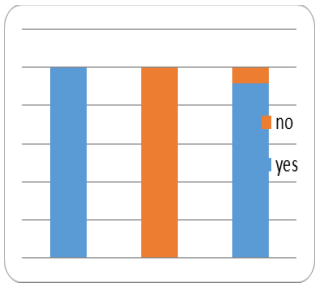
B) Satisfaction with the quality of food



C) Ac room



D) Wifi facility



E) Wifi facility is important

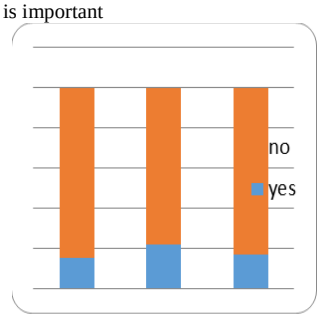


FIGURE 11: Comparison on basis of facilities

It is thus seen that though there is a huge difference in the cost that is paid of the students living in PGs, hostels and flats, the facilities that they receive do no vary much. Though the residents of PG pay more than twice the amount paid by the latter two, the only additional facility that they seem to get is that of AC and of Wifi (if compared to residents of flats).

V. DISCUSSION:

Various studies have been conducted to study the student housing industry in Delhi but none has tried to find the factors that affect the choice of accommodation and the social cost that is incurred in living outside ones native place. This research on the other hand has tried to deeply analyse the factors and the reason why they affect the choice. Additionally, if at all such a study was conducted, it was not solely focused on the students of Delhi University.

This research finds that there is indeed a relation existing between the income level and choice of accommodation. While the study (Gupta et al., 2015) only talked about the fact that only a small percentage of students manage to get hostels in campus, no analysis was made as to the various options available besides in campus facilities and how was the choice affected by income level of the families to which the students belong. Girjia (2018) found about various avenues through which gender discrimination is practised in the society and the way it affects the lives of the girls studying in Delhi Universities. My research goes further in analysing this in view of the curfew time imposed on girls. I have further tried to analyse various other social costs that students pay. None of the studies done before had tried to compare the cost involved in living in PGs, hostels and flats. My study not only tries to compare the cost but also tries to find out what results in such difference and is the difference justified. This is something that is different from all the earlier studies conducted on this topic.

VI. SUMMARY AND CONCLUSION

It is seen that a relation does exist between the income level and choice of accommodation. As the income level rises, the affinity towards PGs increase while that towards flats decrease.

It is seen that there is indeed a relation between gender and the choice of accommodation. Of the respondents living in PGs, the proportion of girls is more. It is further observed that girls give most priority to security which further helps in concluding that they find PGs secure

It is seen that the outstation students face various kinds of discrimination on basis of their religion and the region to which they belong. Though this discrimination is not practised at a large scale, it is something that needs to be dealt with. Besides, this girls face discrimination in regards to the time till which they are allowed to stay out at night.

It is seen that though there is no major difference in the facilities provided, the cost of living in PGs, hostels and flats vary significantly and in case of private PGs it is on an average, more than even than the cost incurred to get the college degree from college.

REFERENCES

[1] Retrieved from url http://www.du.ac.in/du/uploads/rti/act-i.pdf.
Ministry of Human Resource Development. University of Delhi-Act, Statutes and Ordinances. (2004).
[2] Broker, Girija. “Safety First: Perceived Risk of Street Harassment and Educational Choices of Women” (2018).
[3] Batra, Gopal. “Accommodation woes continue to haunt aspiring students in DU”(2018).
[4] Agrawal, Sagar., Gupta, Kounal., & Garg, Mayank. “Scope and comparison of student housing in India.” International journal of Informative and futuristic Research(2015): 2-11.
[5] Garg, Mayank., Gupta, Kunal, Jha, Rakshit. “An Empirical Study on Market Research of Organized Students’ Housing Industry in India”. International Journal of ICT and Management (2014): 2-2.
[6] V. Krishnaiah “Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques” International Journal of Computer Science and Information Technologies, Vol. 4 (1), 39 – 45 www.ijcsit.Com ISSN: 0975-9646, 2013.

[7] Zakaria Suliman zubi “Improves Treatment Programs of Lung Cancer using Data Mining Techniques” Journal of Software Engineering and Applications, 7, 69-77, February 2014.
[8] P Kumar, T Choudhury, S Rawat, S Jayaraman,Analysis of various machine learning algorithms for enhanced opinion mining using twitter data streams,Micro-Electronics and Telecommunication Engineering (ICMETE), 2016.
[9] Anoushka Jain ; Tanupriya Choudhury ; Parveen Mor ; A. Sai Sabitha,Intellectual performance analysis of students by comparing various data mining techniques,3rd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), 2017
[10] T Choudhury, V Kumar, D Nigam, B Mandal, Intelligent classification of lung & oral cancer through diverse data mining algorithms, Micro-Electronics and Telecommunication Engineering (ICMETE), 2016
[11] T Choudhury, V Kumar, D Nigam, An Innovative and Automatic Lung and Oral Cancer Classification Using Soft Computing Techniques,International Journal of Computer Science & Mobile Computing 4 (12), 313-323,2015
[12] D Panchal, P Chatterjee, RK Shukla, T Choudhury, J Tamosaitiene, Integrated Fuzzy Ahp-Codas Framework For Maintenance Decision In Urea Fertilizer Industry, Economic Computation & Economic Cybernetics Studies & Research 51 (3),2017

2018 5th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)

A Multi-layered Outlier Detection Model for Resource Constraint Hierarchical MANET

Adarsh Kumar
School of Computer Science
University of Petroleum and Energy Studies
Dehradun, India
adarsh.kumar@ddn.upes.ac.in

Alok Aggarwal
School of Computer Science
University of Petroleum and Energy Studies
Dehradun, India
alok.aggarwal@ddn.upes.ac.in

Divakar Yadav
Department of Computer Science
M. M. M. University of Technology
Gorakhpur, India
dsycs@mmmut.ac.in

***Abstract*— For sharing resources using ad hoc communication MANET are quite effective and scalable medium. MANET is a distributed, decentralized, dynamic network with no fixed infrastructure, which are self- organized and self-managed. Achieving high security level is a major challenge in case of MANET. Layered architecture is one of the ways for handling security challenges, which enables collection and analysis of data from different security dimensions. This work proposes a novel multi-layered outlier detection algorithm using hierarchical similarity metric with hierarchical categorized data. Network performance with and without the presence of outlier is evaluated for different quality-of-service parameters like percentage of APDR and AT for small (100 to 200 nodes), medium (200 to 1000 nodes) and large (1000 to 3000 nodes) scale networks. For a network with and without outliers minimum improvements observed are 9.1% and 0.61% for APDR and AT respectively while the maximum improvements of 22.1% and 104.1%.**

Keywords— outliers, inliers, attack detection, density-based clustering, QoS.

I. INTRODUCTION

Hierarchical Mobile Ad hoc Networks (MANETs) are resource constraint, autonomous, self-forming & healing networks with very limited hardware resources. These networks are used in various applications like improving coverage extension in future 5G cellular networks, traffic management, disaster management, battlefield secure communication, visitor tracking systems etc. [1]. Security solutions are quite challenging due to limited hardware as well as constraint quality-of-service parameters. Ample research has been done for providing security to these networks [2]. Availability, authentication, authorization, confidentiality, integrity and non-repudiation are major security primitives. Out of all these security primitives, availability of nodes is of the primary concern as availability of nodes as and when required is mandatory to integrate security primitives and services.

An outlier detection mechanism identifies nodes relevant for secure communications with respect to availability of nodes. Statistical approaches are adopted by majority of outlier detection methods either parametric or non-parametric, which uses statistical properties for training and testing [2]. Parametric outlier detection approach, modelled using means and covariance, assumes some underlying distribution such as normal or Gaussian while non-parametric approaches are silent to statistical properties of the data. Univariate and multivariate analysis are also used for identifying outliers. It is observed in [3] that single layer outlier detection solutions are not fruitful, making multi-layer outlier detection solutions a prominent area of research which exploits dependencies between layers which in-turns improves scalability, performance and feature sharing capabilities. Identification of outliers through their performance and features helps in proper identification of misbehaving nodes as each layers suffers from multiple attacks.

This work proposes a multi-layered outlier detection model; consisting of outlier detection at MAC, routing and application layer of MANET stack; for resource constraint hierarchical MANETs. Machine learning and Bayesian classifier model are used for outlier detection at MAC layer. For IP features and transition probability distribution based outlier detection module, routing layer uses training and testing data prepared at MAC layer. For application trust score based feature extraction and rule based outlier detection at the application later, same data is used. An aggregated outlier detection module is also added which identifies outliers based on outlier scores collected from three MANET layers, i.e. MAC, routing and application layers. This later identifies outliers based on outlier scores collected from three MANET layers and identifies outliers at both local group and at hierarchical levels. The proposed approach can identify the outliers using one layer or more hence it gives flexibility to scale the outlier detection complexity based on availability of resources

Rest of the paper is organised as follows. Major works by earlier researchers in multi-layered outlier detection in MANETs are presented in section II. Description of proposed multi-layered outlier detection approach for resource constraint MANETs are given in section III. Results and analysis of simulation are given in Section IV. Finally, section V concludes the work with conclusion.

II. LITERATURE SURVEY

Since MANETs are resource constraint devices hence lighter statistical techniques are often used for the detection of outliers. For these type of networks, various multi-layered models of outlier detection have been proposed [3]-[10]. In [3] authors have proposed a dynamic anomaly detection technique using cross layer for MANET where MAC and routing layers are used for outlier detection with packet drop count and missed IP DSN observations respectively. Proposed approach can detect and isolate black hole attacks from the network.

978-1-5386-5002-8/18/$31.00 ©2018 IEEE

For detecting misbehaving node in MANETs, a two-level outlier detection scheme is proposed in [4] where MAC and network layers are mainly considered for feature selection and outlier detection. A decision tree classification is used for generating instances at first level and accumulated measure of fluctuation of the received classified instances at second level. Concept of variability of the smaller size population is used. Linear regression process is performed for separating the normal nodes from malicious nodes.

Different outlier detection approaches are proposed for MANETs [5]-[10] which mainly uses MAC and routing (network) layers for outlier detection. In [7] application layer is preferred for feature extraction and outlier detection. Machine learning techniques for training or testing is rarely used in the existing single or multi-layered outlier detection techniques and concentrate mainly on data collection and attack detection resulting in the identification of incomplete or inconsistent data records as outliers irrespective of its relevance in completing a useful transaction. Further, existing approaches also do not take into account the complexity of outlier computation, which is an important parameter for resource constraint devices.

III. PROPOSED APPROACH

In this section a multi-layered outlier detection architecture is proposed for identifying misbehaving nodes. The proposed architecture collect misbehaving nodes at different layers of MANET protocol stack. Various modules proposed in this architecture include: data gathering, data pre-processing, MAC layer detection (MACLD), routing layer detection (RLD), application layer detection (ALD) and aggregated layer detection (AGLD). Functionalities of these modules are as follows:

1. *Data Gathering*: In this module, raw data is collected using a pre-defined data collection sensor designed using threshold based QoS parameters. In raw data, attributes of nodes and their neighbouring nodes are collected for analysis.

2. *Data Pre-processing*: The collected data is analysed for removing duplicate entries, handling missing entities, data enrichment using node attributes etc. The pre-processed data is analysed for re-presentation of information as desired in analysis process.

3. *MAC layer detection (MACLD):* This module applies machine learning mechanism for outlier detection over MAC layer packets. Four phases of this module are: pre-processing, learning/training, evaluation and prediction. The pre-processing phase prepared training and test data sets from raw data received from data pre-processing module. Trained dataset includes data labels for each node as outlier or inlier. Features of test data is compared with training dataset for new labels. Features include association of source and destination nodes, signal strength, node's QoS performance, packet prioritization mechanism, route overriding mechanism, reliability and recoverability of node and, route finding capability and fault tolerate strength. Out of content or context aware representation, node profile is context aware representation of its features. Node is preferred to be considered as inlier if training dataset results in preparing a node profile with score greater than average value of all node values prepared using divisive hierarchical clustering algorithm [21]. Evaluation phase of this module uses test dataset for extracting testing node's features, prepare node profile and apply Adaptive k-Dependency Bayesian Contextual Outlier Detection (Adaptive k-DBCOD) process for outlier detection. Pseudocode 1 explains Adaptive k-DBCOD process in detail.

Pseudocode 1: Adaptive k-Dependency Bayesian Contextual Outlier Detection (Adaptive k-DBCOD)
Goal: To predict the class of each data point in dataset with maximum accuracy.

1. Pick a node for evaluation and extract its features using predefined format.
2. Apply Naïve Bayes structure and determine whether extracted feature has a maximum of *k*-feature nodes as its immediate parent node.
3. If feature similarity score of a node and its immediate parent node is greater than a certain threshold value then
    a. Apply univariate gaussian predictor over extracted feature to test content based outlier.
    b. If content based outlier is detected then complete node profile is extracted for context based multivariate gaussian predictor.
    c. If context based outlier is detected then results is returned as outlier.

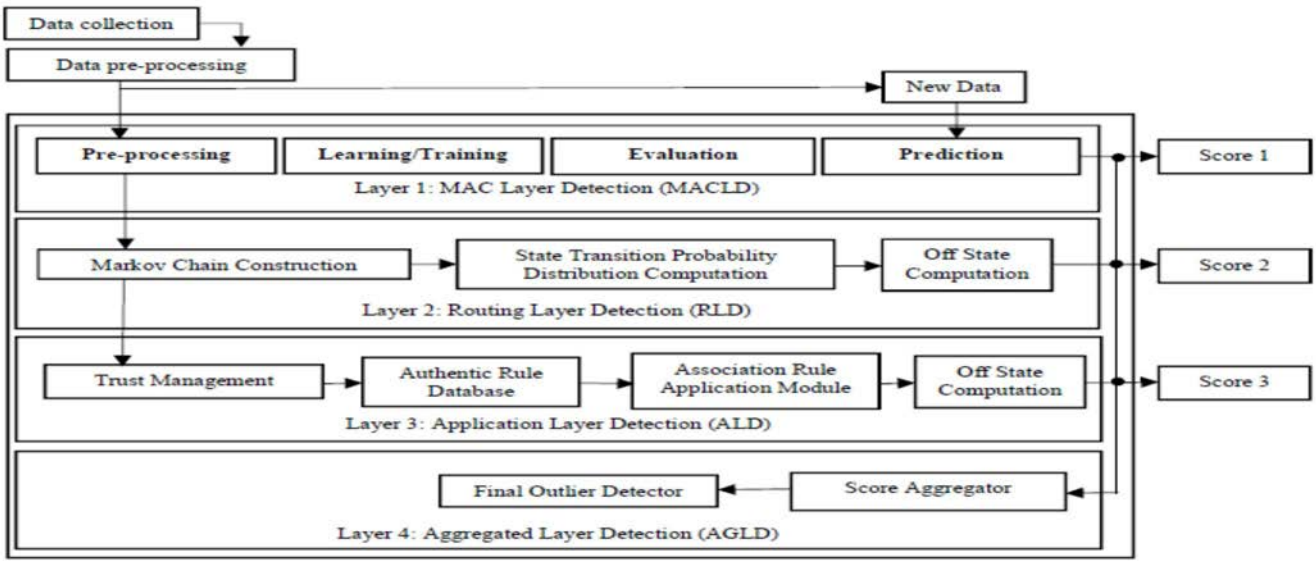


Figure 1: Proposed Multi-layered Outlier Detection Architecture

d. Else-if context based outlier fails then result is returned as inlier.

4. If feature similarity score of node and its immediate parent node is lesser than a certain threshold value then node is considered as inlier.

Pseudocode 1 applies univariate gaussian predictor function and multivariate gaussian predictor for outlier detection process. Univariate gaussian predictor function evaluates the node features against historical data of same node. This function predict the value using historical data and compare the result with value computed after testing. This prediction is truly based on content based outlier. Whereas, multivariate gaussian predictor is both content and context based outlier detection mechanism. Node profile is prepared before applying multivariate gaussian predictor function. Node profile keep records of feature extracted for testing node as well as nodes with similar features in the network. Multivariate gaussian predictor function compares node features similarity score with average value of all nodes found in node profile. Fig. 2 shows how k varies in Adaptive k-DBCOD process. Fig. 2(a) shows a directed acyclic graph of three nodes under outlier detection process. In this example, every node is individual and no node profile matches with neighboring or parent node thus k=0. Figure 2(b) shows an example of DAG where maximum number of parent nodes available for any node is one. Thus, features of node 2 and node 3 are compared with feature of node 1 for similarity check. Fig. 2(c) shows an example of DAG where maximum number of parent nodes available for any node is two. Thus, features of node 3 are compared with feature of node 1 and node 2 for similarity check.

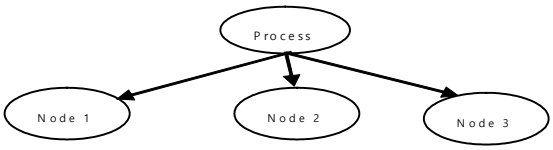


FIGURE 2(a): Directed Acyclic Graph (DAG) when k=0

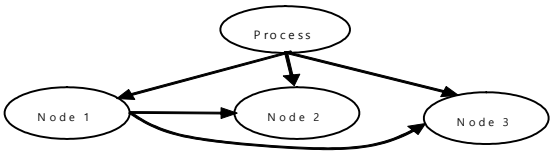


FIGURE 2(b): Directed Acyclic Graph (DAG) when k=1

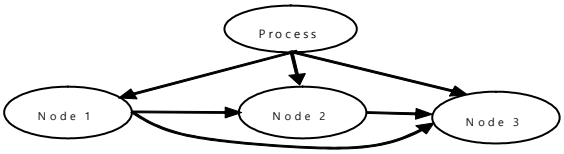


FIGURE 2(c): Directed Acyclic Graph (DAG) when k=2

Figure 2: Examples of variations in k for Directed Acyclic Graphs (DAGs)

If a node passes content and context based outlier detection test then it is considered as outlier else inlier. A parallel process of random context check may also be executed in order to save time or speed up the detection process. Further, random context based check method randomly predict the inlier for comparison in opposite to outlier. Final label of evaluation phase is marked when outcome of Adaptive k-DBCOD process is compared with training phase. If results of both process is same then test data point is marked with outlier and outcome is appended in training dataset. Finally, prediction phase also speedup the outlier detection by building the node profile of every new node and compared its features with existing node profiles. A MAC layer scorer (MACLS) phase calculates percentage of outliers in a particular cluster or set of clusters under testing. Outlier detection process implementer may use MACLS value as final value of outlier detection process and stop further outlier detection computations, or move to routing layer for comprehensive and increasing computational cost based outlier detection. Pseudocode 1 is based on both content and context based outlier detection process defined in [8]. Mathematically, probability of content based outlier detection is computed as:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Where, $\mu = \frac{1}{m}\sum_{i=1}^{m} x^{(i)}$ and $\sigma^2 = \frac{1}{m}\sum_{i=1}^{m}(x^{(i)} - \mu)^2$

Further, probability of context based outlier detection is computed as follows:

$$p(x) = \frac{1}{(2\pi)^{(\frac{n}{2})}|\sum|^{(\frac{1}{2})}} e^{(-\frac{1}{2}(x-\mu)^T \sum^{-1}(x-\mu))}$$

Where, $\mu = \frac{1}{m}\sum_{i=1}^{m} x^{(i)}$ and $\sum = \frac{1}{m}\sum_{i=1}^{m}(x^{(i)} - \mu)(x^{(i)} - \mu)^T$

4. *Routing layer detection (RLD):* In this module, outliers are detected by predicting the state of a node and constructing a Markov chain. This Markov chain represents the transition between certain states. In MANETs, Markov chain draws a sequence of node's transition between states without referring to any historical record. After drawing state transition diagram, two types of control packets are observed: request-to-send (RTS) and control-to-send (CTS). If any node, in any of its state, send or receive RTS or CTS packets more than a certain threshold then those nodes are observed for outlier detection. In RLD, the whole process of outlier detection is divided into three modules: constructing Markov chain, N-Gram transitional probability distribution based outlier detection and routing layer scorer (RLS). *Markov chain construction* is further divided into three sequential components: generate probability transition matrix, plot transition flow and compute steady-state vector. A probability transition matrix is computed as follows:

$$\mathrm{P}_{m\,x\,n} = [[P_{0,0}, P_{0,1}, \ldots . P_{0,n}], [P_{1,0}, P_{1,1}, \ldots . P_{1,n}], \ldots, [P_{m,0}, P_{m,1}, \ldots . P_{m,n}]]$$

Where, P_{ij} represents the probability of a node to transit from i^{th} state to j^{th} state.

This probability transition matrix is helpful in plotting transition flow. For example, a three state Markov chain transition plot is shown in fig. 3. In this example, there are three states: listener, sender and receiver. $P_{Listner,Listner}$, $P_{Listner,Sender}$, $P_{Sender,Sender}$, $P_{Sender,Listner}$, $P_{Sender,Receiver}$, $P_{Receiver,Sender}$, $P_{Sender,Listner}$, $P_{Receiver,Receiver}$ and $P_{Listner,Sender}$ are the probabilities of various transitions among listener, receiver and sender states. Further, if $\mathrm{P}_{m\,x\,n} . \lambda = \lambda$ is satisfied then λ is considered as steady-state vector of a regular Markov chain. Second phase

of RLD is N-Gram transitional probability distribution based outlier detection. This functionality is explained in pseudocode 2.

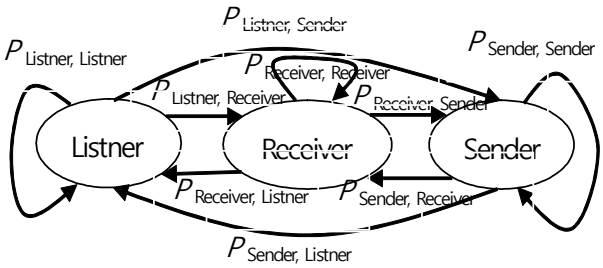


FIGURE 3: Three state Markov chain transition plot

Pseudocode 2: *N*Gram - transitional probability distribution based outlier detection

1. Initially, every node constructs a *N*-gram Markov model representing a sequence of states.
2. State probability transition matrix is pre-computed, value of probability transition is extracted and assigned to N-gram Markov model.
3. A recurrence function of sub-sequence is generated with length of *N*-grams. Here, sub-sequence is the part of state sequences followed by every node.
4. Two sequences are compared for feature similarity: first sequence is as computed in step 3 and second sequence is extracted from training dataset with similar state pattern.
5. Two sequences generated in step 4 are passed to odds ratio test i.e. ratio of probability of sequence computed in step 3 and sequence extracted from training dataset with similar state pattern is calculated. If result of this ratio is greater than one then nodes following sequence are considered as inliers else outliers.
6. All sequences passing the odds ratio test are passed to threshold based tests. This test checks the state of a node. If it is following a state sequence of listener, receiver, forwarded, idle and sleep, and number of request to send packets are greater than a certain threshold then node is considered as outlier. Similarly, if node is following a state sequence of sender, forwarded, idle and sleep, and number of control to send packets are less than a certain threshold then node is considered as outlier else inlier.

RLS phase calculates percentage of outliers in a particular cluster or set of clusters under testing. Outlier detection process implementer may use RLS value as final value of outlier detection process and stop further outlier detection computations, or move to application layer for comprehensive and increasing computational cost based outlier detection.

5. *Application layer detection (ALD):* In this module, rule mining is applied for outlier detection. Mining association rules [9][10] and fuzzy logic [13] are preferred method for attack detection through outlier detection mechanisms in hierarchical networks. In [11], similar mechanism is proposed using fuzzy logic based imperialist competitive clustering algorithm. In this algorithm, cluster head decides its cluster density through fuzzy rules and it varies over time. The proposed approach for outlier detection at application layer is modified form of imperialist competitive clustering algorithm. The proposed approach starts with randomly deployment of nodes (population) and a set of closely linked clusters are considered as a single colony for analysis. Further, a set of colonies constitutes an empire. Details of proposed imperialist competitive clustering based algorithm is presented in pseudocode 3.

Pseudocode 3: Imperialist competitive clustering with reduced cost and power consumptions

1. Apply divisive hierarchical clustering algorithm [21] and generates initial population with reduced noise.
2. Generate population matrix, identify a data point with minimum cost and compute its distance from all other data points considered in its colony. Average distance of data point under evaluation is its cost.
3. Initial population after reducing noise is sorted on cost values.
4. Two imperialist zones are created: imperialist states and imperialist countries. Imperialist states consists of colonies with maximum cost values and remaining colonies are assigned to imperialist countries.
5. Compute normalized cost and normalized power of imperialists after dividing colonies into imperialist states and countries.
6. Normalized cost of any colony is computed by subtracting the cost of colony from maximum value of cost from all colonies. Normalized power is computed by dividing normalized cost with sum of costs from all colonies. For further analysis, initial set of colonies are decided based on certain threshold limit for normalized power.
7. An empire is build from colonies using divisive hierarchical clustering algorithm.
8. Select each empire one by one, extract every colony of it and move the colony towards imperialist state using fuzzy min-max. Cost value of two populations (new colonies and old colonies) are compared for inlier and outlier detection. Those colonies are considered as inliers whose cost value is maximum.
9. Soft all colonies again based on its cost values and select top-X colonies with minimum power for outlier detection.
10. Calculate cost of all empires and identify weakest empires. Move weakest empires to its neighboring best cost empire using node's mutation features.
11. Those empires are not able to move, are considered in outlier categories.

ALS phase calculates percentage of outliers in a particular cluster or set of clusters under testing. Outlier detection process implementer may use ALS value as final value of outlier detection process and stop further outlier detection computations, or move to aggregated layer for comprehensive and increasing computational cost based outlier detection.

7 *Aggregated layer detection (AGLD):* In this module, information about each node is collected individually from MACLD, RLD and ALD. If percentage of outliers computed from each node's label and percentage of outliers collected from above three layers is same then this value is considered as final and value and no further processing is being done else

fuzzy min-max mechanism is applied for final score calculations. Details of minim-max mechanism application is presented in pseudocode 4.

Pseudocode 4: Aggregated Layer Outlier Score Calculator

1. Collect each node's label and percentage of outlier values from above three layers.
2. Now, if percentage of outlier calculated from above three layers and percentage of outlier value collected from above three layers are different then apply divisive hierarchical clustering algorithm and create clusters for computing outliers in local groups (colonies) and global groups (empires).
3. Percentage and label values computed from colonies and empires are considered as final outcome of proposed outlier detection approach.

IV. Results and Analysis

This section starts with explanation to simulation parameters considered for analysis. Further, internal and external indices are used for validating the colonies and empires. Along with internal and external indices, performance and QoS parameters are evaluated for outlier detection. Details of each of said things are explained in following subsections:

A. Simulation Setup

Simulation parameters used in this work are shown in table 1. Simulation is performed over a minimum of 100 and maximum of 3000 nodes. Mobility of nodes varies from 0.5 m/s to 7 m/s randomly in an obstacle free area of 1000x1000 sq. meters.

Table 1: Simulation Parameters

Parameters	Value
Number of nodes	100 to 3000
Channel Type	WirelessChannel
Radio Propagation Model	Ray Tracing
Network Interface	WirelessPhy
MAC Type	802.11
Interface Queue	Priority Queue
Antenna	OmniAntenna
Max Packets in Queue	50
XxY dimensions of the topography	1000x1000 sq. meters
Mobility Model	Random WayPoint
Data Rates	5 packets/second
Packet Size	512 bits
Simulator	ns-3[12]
Simulation Time	1000sec
Number of slots assigned to reader at stretch (Δ)	1
Time of each slot	10 msec.
Velocity (Minimum to Maximum)	0.5 m/s to 7 m/s

B. Analysis of Internal and External Cluster Indices

In order to measure the quality of clusters formed, internal and external indices are measured. Quality of clusters increases with clustering indices value improve. This process is helpful in identifying outliers and inliers among clusters and in whole network. In this experimentation, total simulation time is divided into thirteen slots and index value is computed in each slots. Details analysis of index calculations are presented in following sections:

- Analysis of Internal Indices

Internal indices use working dataset and its inherited properties for measurement rather referring information from external resources or datasets. Two of the popular internal indices used for analysis in this work are: Dunn Index (DI) and Root-Mean-Square Std. Dev. Index (RMSSDI). Results of both of these indices are shown in fig. 4. DI and RMSSDI indices values are increasing for small scale network (100 to 200 nodes) and decreasing for large scale network (200 to 3000 nodes). In case of DI, clusters are considered to be stable if its value increases with increase in participants. In conclusion, probability of clusters stability is increasing in proposed scenario till T9-T10 time slot. Thereafter, it is either constant or decreases with increase in nodes. RMSSDI is another internal index. If RMSSDI value increases slowly in initial duration and exponentially after certain duration then clusters are considered stable. Proposed scenario is giving similar results for almost all number of nodes. Thus, clusters formed in proposed scenarios are acceptable for further outlier detection process.

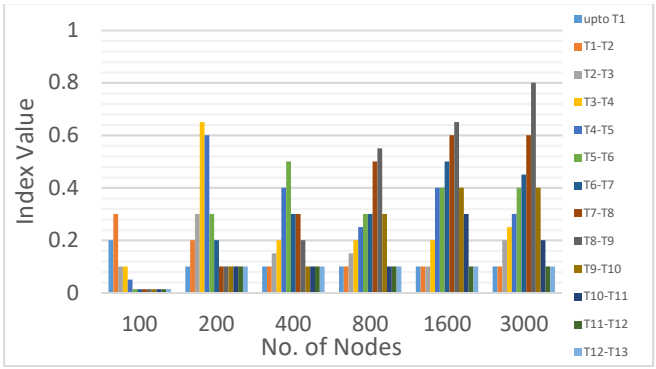


Figure 4(a): Dunn Index (DI)

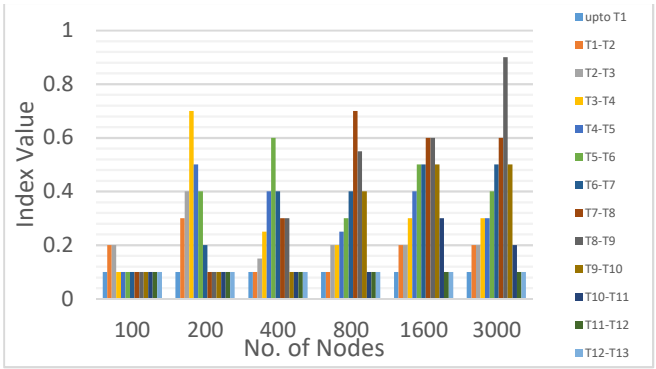


Figure 4(b): Root-mean-square std dev Index (RMSSDI)

Figure 4: Internal Cluster Evaluation

- Analysis of External Cluster Indices

In external indices measurements, quality of clusters is put in acceptable category after considering external database or information i.e. quantities and features inherited from external known sources. Two external indices used for measurements in this work are: G-measure index (FI) and Normalised Mutual Information (NMI). Results of these two indices are shown in fig. 5. In proposed network scenario, number of

nodes are varying from 100 to 3000. In proposed work, a network is considered as small scale network if number of nodes are varying from 100 to 200 and it is considered as large scale network if number of nodes are varying from 200 to 3000. Results show that FI and NMII indices are increasing for 100 to 200 nodes and decreasing for 200 to 500 nodes. Clusters in external indices measurements are considered in stable category if their values are increasing constantly. In case of FI, proposed mechanism is considered in acceptable category if number of nodes in a network are 100, 200 or 800. In other networks, FI value can be considered in acceptable category during initial timing slots i.e. upto T5 time slot as shown in fig. 5(a). Fig. 5(b) shows NMII value variation and results show that proposed mechanism perform better for 100 and 200 nodes network as compared to other networks. Other network shows higher variations that makes clusters in unacceptable category.

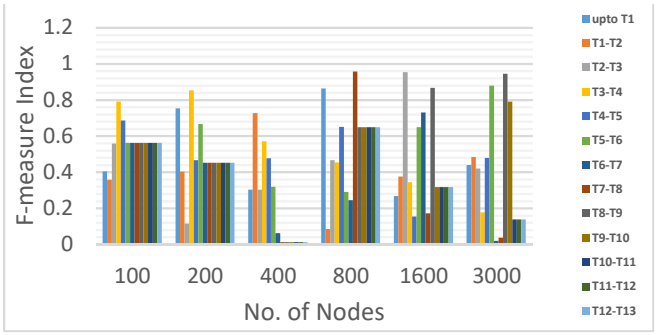


Figure 5(a): F-measure Index (FI)

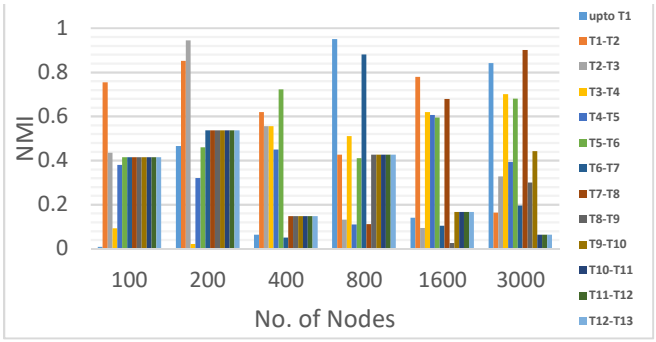


Figure 5(b): NMI

Figure 5: External Cluster Evaluation

C. Analysis of Cluster Performance Indices

In cluster performance indices, clusters are considered in acceptable category by observing the performance behaviours of cluster members, cluster heads and their overall performance. In order to measure cluster performance indices, average cluster head duration (ACD) and average cluster member duration (ACMD) are used in this work. ACD is measure as ratio of average length of time when a node is unanimously selected as cluster-head. If a node is selected as cluster-head for a long duration then it is considered as most acceptable scenario because the selected node is trained over time with proper behaviours of cluster, its nodes and cluster's network. Fig. 6 shows the comparative analysis of ACD for proposed approach with other approaches like: Low-Energy Localized Clustering for HOMOgenous RFID Networks (HOMO-LLCR) [13], Transmission and Collusion Aware Clustering with enhanced Weight Clustering Algorithm (TCACWCA) [14], TCACWCA Expected Transmission Count (TCACWCAETX) [14], TCACWCA Path Encounter Rate (TCACWCAPER) [14], k-hop Compound Metric Based Clustering (KCMBC) [15], Max-Min heuristic approach[16], CBKC scheme [17] based on lowest ID (CBKC_ID) and CBKC scheme based on highest degree (CBKC_Degree). Results show that the proposed approach is better compared to other approaches in terms of percentage increase in ACD value for 1600 to 3000 nodes because frequency of change of cluster head in large network is comparatively lesser.

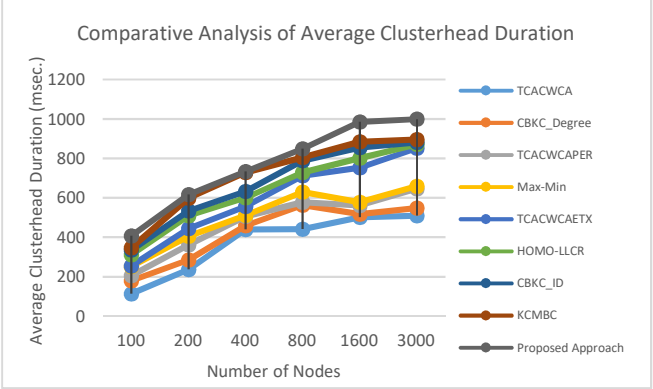


Figure 6(a): Comparative analysis of Average Clusterhead Duration (ACD) for proposed approach with other approaches (velocity = 0.5 m/s to 7 m/s and maximum hops between cluster-head and cluster member = 4)

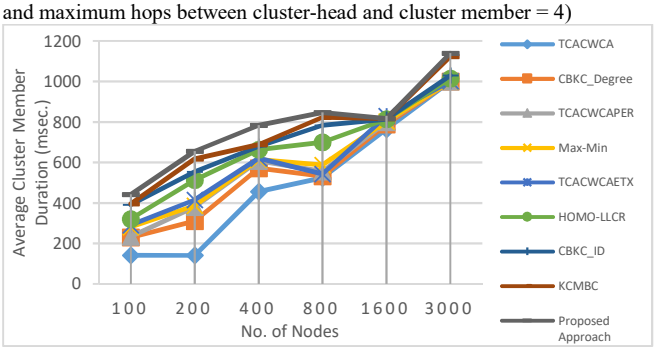


Figure 6(b): Comparative Analysis of Average Clusterhead Member Duration (ACMD) for proposed approach with other approaches (velocity = 0.5 m/s to 7 m/s and maximum hops between cluster-head and cluster member = 4)

Figure 6: Performance Analysis of Cluster Performance Indices

ACMD is another performance indices used for evaluating the clusters. It is the average length of time when nodes remains members of a cluster without changing the cluster-head as well. Fig. 6(b) shows the comparative analysis of proposed approach with HOMO-LLCR, TCACWCA, TCACWCAETX, TCACWCAPER, KCMBC, Max-Min, CBKC_ID and CBKC Degree. It is observed that the proposed approach is comparable to existing approaches with variation in number of nodes over time. Percentage increase in ACMD is better when number of nodes varies from 1600 to 3000 as compared to variation in number of nodes from 100 to 1600. It is because cluster head stability and cluster member stationary value are better for large scale network compared to small scale network.

D. Analysis of QoS Parameters

In order to measure QoS parameters, Average Packet Delivery Rate (APDR) and Average Throughput (AT) are the parameters taken for analysis. APDR is measured as the ratio of number of packets received by destination to the total number of packets sent by the source over a given period of simulation time. Fig. 7(a) shows the comparative analysis of APDR with and without presence of outliers when number of

nodes varies from 100 to 3000. If outliers are present in the network then variation in APDR is below minimum threshold value considered for outlier detection. It is also observed that a minimum of 9.1% (for 1600 nodes) and maximum of 22.1% (for 3000 nodes) improvement is observed in APDR when it is measured without presence of outliers as compared to presence of outliers. Further, AT is defined as the ratio of total size of packets received by destination nodes divided by simulation time. Detailed analysis of AT for nodes variations from 100 to 3000 is shown in fig. 7(b). This is a comparative analysis of AT with variations in number of nodes for two scenarios: with and without presence of outliers. Fig. 7(b) shows that variation in AT value is below threshold value variation decided for outlier detection in this work. When the number of nodes are varying from 100 to 400 or 1600 to 3000 then a minimum difference between AT values of threshold and presence of outliers is computed. A minimum of 0.61% improvement is observed in overall analysis for 1600 nodes when two scenarios (with and without presence of outliers) are compared. Similarly, a maximum of 104.1% improvement is observed for 800 nodes when two scenarios (with and without presence of outliers) are compared.

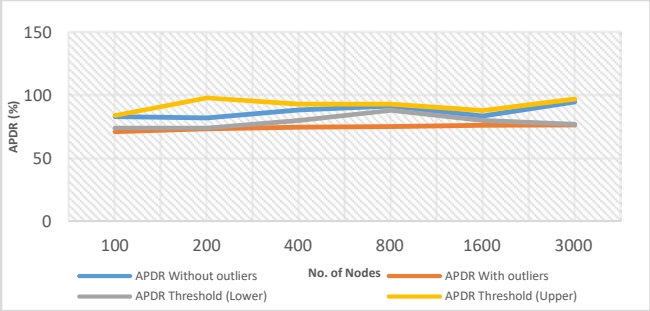


Figure 7(a): Comparative analysis of APDR with variations in number of nodes

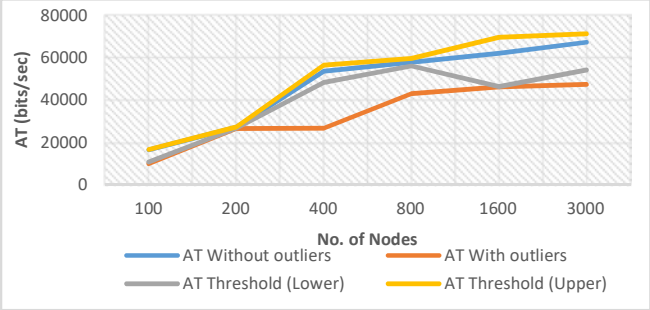


Figure 7(b): Comparative analysis of AT with variations in number of nodes

Figure 7: Performance Analysis of QoS Parameters

V. Conclusion

In order to given efficient and effective security to MANETs, multidimensional-multilayer solutions are required. In this work, availability cryptography primitive is ensured through outlier detection approach for hierarchical MANETs. In proposed hierarchical MANETs, data attributes are also divided hierarchically for efficient distance measurement metric. In proposed outlier detection approach, outliers are identified using MACLD, RLD and ALD modules. Further, AGLD module collects percentage of outlier values and outlier labels from MACLS, RLD and ALD for computing an aggregated decision. AGLD module provides analysis at both local (cluster) and global (network) level in its output. In simulation analysis, indices and performance analysis is performed for measuring the stability of network. Indices values are classified as: internal, external and performance. Results show that network is stable for 100 to 3000 nodes with proposed outlier detection approach. Further, QoS parameters used for performance analysis are APDR and AT. Results show that the proposed approach gives 9.1% to 22.1% improvement in APDR for a network of 100 to 3000 nodes. Similarly an improvement of 0.61% to 104.1% in AT is observed for network without outliers using proposed approach.

References

[1] J. Liu, Y. Xu, Y. Shen, X. Jiang, and T. Taleb, "On Performance Modeling for MANETs Under General Limited Buffer Constraint," *IEEE Trans. Veh. Technol.*, vol. 66, no. 10, pp. 9483–9497, Oct. 2017.

[2] S. Sen, J. A. Clark, and J. E. Tapiador, *Security Threats in Mobile Ad hoc Netwroks*. 2016.

[3] G. Usha, M. Rajesh Babu, and S. S. Kumar, "Dynamic anomaly detection using cross layer security in MANET," *Comput. Electr. Eng.*, vol. 59, pp. 231–241, Apr. 2017.

[4] A. Amouri, S. Morgera, M. Bencherif, and R. Manthena, "A Cross-Layer, Anomaly-Based IDS for WSN and MANET," *Sensors*, vol. 18, no. 2, p. 651, Feb. 2018.

[5] I. Butun, S. D. Morgera, and R. Sankar, "A Survey of Intrusion Detection Systems in Wireless Sensor Networks," *IEEE Commun. Surv. Tutorials*, vol. 16, no. 1, pp. 266–282, 2014.

[6] L. Nishani and M. Biba, "Machine learning for intrusion detection in MANET : a state-of-the-art survey," *J. Intell. Inf. Syst.*, 2015.

[7] A. Amouri, V. T. Alaparthy, and S. D. Morgera, "Cross layer-based intrusion detection based on network behavior for IoT," in *2018 IEEE 19th Wireless and Microwave Technology Conference (WAMICON)*, 2018, pp. 1–4.

[8] M. A. Hayes and M. A. Capretz, "Contextual anomaly detection framework for big sensor data," *J. Big Data*, 2015.

[9] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," *ACM SIGMOD Rec.*, 1993.

[10] M. Hahsler and R. Karpienko, "Visualizing association rules in hierarchical groups," *J. Bus. Econ.*, 2017.

[11] S. Shamshirband, A. Amini, N. B. Anuar, M. L. Mat Kiah, Y. W. Teh, and S. Furnell, "D-FICCA: A density-based fuzzy imperialist competitive clustering algorithm for intrusion detection in wireless sensor networks," *Meas. J. Int. Meas. Confed.*, vol. 55, pp. 212–226, 2014.

[12] "The Network Simulator - ns-2." [Online]. Available: https://www.isi.edu/nsnam/ns/. [Accessed: 05-Jul-2018].

[13] J. Kim, W. Lee, J. Yu, J. Myung, E. Kim, and C. Lee, "Effect of localized optimal clustering for reader anti-collision in RFID networks: Fairness aspects to the readers," *Proc. - Int. Conf. Comput. Commun. Networks, ICCCN*, vol. 2005, pp. 497–502, 2005.

[14] T. Maragatham, S. Karthik, and R. M. Bhavadharini, "TCACWCA: transmission and collusion aware clustering with enhanced weight clustering algorithm for mobile ad hoc networks," *Cluster Comput.*, 2018.

[15] S. Leng, Y. Zhang, H.-H. Chen, L. Zhang, and K. Liu, "A Novel k-Hop Compound Metric Based Clustering Scheme for Ad Hoc Wireless Networks," *IEEE Trans. Wirel. Commun.*, vol. 8, no. 1, 2009.

[16] A. D. Amis, R. Prakash, T. H. P. Vuong, and D. T. Huynh, "Max-min d-cluster formation in wireless ad hoc networks," in *Proceedings IEEE INFOCOM 2000. Conference on Computer Communications. Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies (Cat. No.00CH37064)*.

[17] G. Chen, F. G. Nocetti, J. S. Gonzalez, and I. Stojmenovic, "Connectivity based k-hop clustering in wireless networks," in *Proceedings of the Annual Hawaii International Conference on System Sciences*, 2002.

An Intelligent SLA based Cloudlet Allocation Strategy using Machine Learning

Vinayak Bajoria
Department of Analytics
University of Petroleum and Energy Studies
Dehradun, India
bajoriavinayak@gmail.com

Yash Agrawal
Department of Analytics
University of Petroleum and Energy Studies
Dehradun, India
yash.yash.agarwal1997@gmail.com

Avita Katal
Department of Virtualization
University of Petroleum and Energy Studies
Dehradun, India
avita207@gmail.com

***Abstract*—The internet has widened its horizon far beyond the web browser. The service oriented technologies like Live Streaming videos over Netflix, Hulu, YouTube make up to 60% of today's internet traffic. These services require massive computation power; cloud computing which is onset conveyance of computational power, database storage, and other IT related services through a cloud platform via internet on pay-as-you-go pricing model serves the required purpose. However, the computational power available at the data centers is often limited and should be utilized efficiently so that the multiple request can be handled simultaneously and the terms and conditions reached between the client and the Cloud Service Provider are honored. In this paper, Multilevel Feedback Queue is used for load balancing followed by an SLA based resource allocation i.e. Virtual Machines to the requests i.e. cloudlets using the A* algorithm. The A* derives its values from the Ant Colony Optimization and Heuristic search. Together they form a Machine Learning technique that eventually learns the path that should be followed by a particular size of cloudlets. Thus, decreases the execution time of the cloudlets while necessarily maintaining the Quality of Service.**

Keywords—Artificial Intelligence; A*; Multilevel Feedback Queue; Vague Theory; Heuristic; Ant Colony; pheromone; Cloudlet; Optimization; Quality of Service(QoS)

I. INTRODUCTION

Cloud Computing is on demand delivery of computing power, data base storage and other IT related services through a cloud service platform over the internet with pay-as-you-go pricing scheme. These service request can be of any form like live streaming videos, backing up personal photos to the cloud storage, request to launch certain kind of business related service like Net Banking, Payment Gateways. These entire requests are in form of an instruction set known as cloudlet that has to be computationally executed upon a server which deploys a Java Virtual Machine instance to execute the request.

This paper focuses to increase the Quality of Service provided to the client by decreasing the total execution time of the cloudlets while keeping in mind the priority set by the user for the cloudlet which makes the cloudlet either time-efficient or cost-efficient.

II. RELATED WORK

Many researchers have made significant contributions towards solving the NP hard problem of cloudlet allocation.

In Scheduling algorithms like First Come First Service [2] cloudlets are simply allocated to the first available Virtual Machine. The strategy has least complexity but has a leads to starvation of cloudlet especially of smaller cloudlets. Shortest Job First [2] strategy sorts the cloudlets in ascending order before applying the FCFS. However, the strategy is only an improvement for the cloudlets with shorter lengths. Hadi Gougrazi [3] proposes a resource allocation problem aiming to decrease the power consumption of the whole cloud computing system while fulfilling the SLA in probabilistic sense. A penalty is issued if SLA violation occurs. A heuristic algorithm is used to solve the resource allocation problem. Vinayak et. al. [4] proposes a cloudlet allocation strategy that uses multilevel feedback queues which run cloudlets for a specified time quantum which removes the need of load balancing. However, the algorithm seems to run only better in comparison to FCFS effectively. Nishant et. al. [5] proposed an Ant Colony Optimization (ACO) technique for efficient scheduling of the incoming tasks by making use of the underutilized nodes in the cloud environment. However, it has been observed that the designed algorithm after certain iteration slows down and is unable to change ths status of the node for future scheduling of tasks. Reza Shojaee et. al. [6] proposed a Cat Swarm Optimization for scheduling tasks in distributed environment. However, the algorithms seek and trace the modes of CSO which has lead to delay in the execution time. Shridhar G. Domanal and G Ram Mohana Reddy [7] [8] proposed a modified throttled algorithm that efficiently schedules the incoming client tasks to the virtual machines. The authors focus was on the response time of the tasks. Glauco Goncalves et al. [9] proposed a Distributed Cloud Resource Allocation System(D-CRAS)which ensures an automatic monitoring and control of resources of the cloud to guarantee the optimal functioning while meeting the SLA requirements. Wenyu Zhou

et al. [10] implemented a resource allocation policy for load balancing in virtual machine cluster. This resource allocation policy not only monitors the real-time resource utilization of CPU, memory etc. but also uses instant resource reallocation for VMs running on the PM using VM migration.

III. PROPOSED WORK

In this paper, we derive our inspiration from the Travelling Salesman Problem and Ant Colonies. The cloudlets enter a Multilevel Feedback Queue where, the time quantum for the queue is calculated. Then these cloudlets are categorized into batches and sorted. Each batch enters the Machine learning model where a cloudlet acts like a salesman and executes on every Virtual Machine for the time quantum calculated earlier. The order in which the cloudlets visits the cities i.e. Virtual Machines is its path. To find this path and make it optimal eventually, A* algorithm is used. A* algorithm uses the Ant Colony optimization technique to find the partial solution and Heuristic function to make the search move towards the goal i.e. is to completely execute the cloudlets. Once the cloudlet completes a tour and if the instructions are still left to be executed, it moves to a lower priority queue.

A. Data Base

Before moving on with the proposed work we would like to set some naming conventions which will be used throughout the paper.

1) Cloudlet: Each request or task or salesman or cloudlet can be expressed as C (id, a, l, e, w, rvv, vv, t, f, b) where 'id' identifies each request uniquely, 'a' is the arrival time of the cloudlet, 'l' is the instructions in the cloudlet on the scale of millions, 'e' is the execution time of the cloudlet which is 0 at starting, 'w' is the waiting time. 'rvv' is the recent VM visited by the cloudlet in a particular tour of the model. It is set to null whenever the cloudlet enters the model. 'vv' is the set of Virtual Machines visited, it is the path followed and 't' is the time quantum for which the cloudlet will run on particular VM when it enters the model and 'f' is a flag which can have 0 or 1 value. 'b' is the batch to which the particular cloudlet belongs.

2) Virtual Machine: Each Virtual Machine can be expressed as VM(id, m, s, pe) where 'id' identifies each virtual machine uniquely, 'm' is the processing power of the virtual machine, 's' is the status of the Virtual Machine which depicts if it is processing some cloudlet or not. 'pe' is the processing elements of the Virtual Machine.

3) Model: To The model is designed by taking inspiration from the Travelling Salesman. The cloudlet is treated like a salesman who has to visit every city i.e. the Virtual Machine only once. The nodes in a particular iteration of the graph are like cities that a salesman can visit. The salesman can visit only one city in a particular iteration.

Each cloudlet can execute for a maximum time period equivalent to the time quantum of the queue at every city i.e. Virtual Machine it visits. The time constraint specified solves the problem of unbalanced loads at the virtual machines. The formulation of the problem as TSP also gives each cloudlet a chance to run on every virtual machine at least once.

The model is a connected NVM*NVM graph where NVM is the number of Virtual Machines. Each column 0 to NVM-1 represents iteration. There are NVM number of Virtual Machines in iteration. The nodes in certain iteration are connected to all the nodes in the next iteration. At each node there exists a different Virtual Machine. The connections in the graph represent the pheromone trail between two nodes.

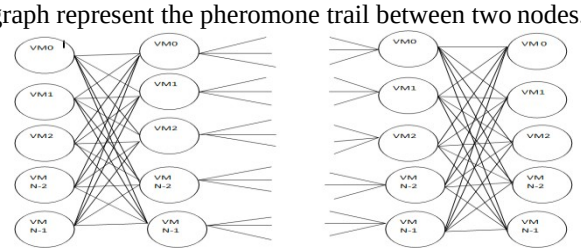


Fig.1. The overview of the model

B. Multilevel Feedback Queue

The strategy makes use of a multiple feedback queues to maintain the cloudlets. For each of the queue the time quantum is calculated on the basis of the vague theory. Time quantum is the amount of the time that each cloudlet can run for upon a particular Virtual Machine (VM) in the model at any iteration.

$Q = \{ Q_1, Q_2, Q_3 ..... Q_m\}$ where Q_i is a multilevel feedback queue formed dynamically.

$L_{max} = \max\{C(id, l)\}$ where $1 \leq id \leq N$, N is the number of Cloudlets and C(id,l) returns the length of cloudlet with certain id. L_{max} is the maximum size of all the cloudlets present. Similarly we can find L_{min}, and L_{avg}.

The static cost quantum Q_s is a value set by the CSP depicting how the various services are charged. The CSP can increase and decrease the cost quantum to change the pricing scheme for the services provided.

Q_s=(CostPerInstruction+CostPerMemoryAccess+CostPerStorage+CostPerBandwidth) (1)

1) Inference System: The algorithm has the ability to learn from itself, it learns from the currently active cloudlets and uses the set of inputs to convert them to a set of desired outputs. The aim is to generate an optimum time quantum for every queue. It has 4 modules

a) Logic Module: Convert the inputs into the respective vague values to handle the dependability of the tasks in universe of disclosure [0,1]. These vague values are defined with true membership function (t_q) and false membership function (f_q). They are calculated as below:

$$t_q = (L_{avg})/(L_{avg} + L_{max} + N) \quad (2)$$

$$fq = (L_{avg})/(L_{avg} + L_{min} + N) \quad (3)$$

b) Grade Module: It defines the degree of accuracy of the vague if it lies within the universe of [0,1].

$$S_q = t_q + f_q \quad (4)$$

c) Inference Module: This returns the optimum value of time quantum. It fetches the value of Q_s from the database and the degree of accuracy from the grade module. Finally, on the basis of the given rules it returns the size of time quantum Q_d.

If (for i=1.....N | A_i = = 0) then

$$Q_d = S_q * Q_s \quad (5)$$

else

$$Q_d = Q_s \quad (6)$$

Where Q_s, is a static Cost Quantum whose value is assigned by the Cloud Service Provider (CSP).

d) Dynamic Module: The time quantum for the next lower priority queue is calculated by a formulae which takes the time quantum of Q_i and the currently number of active cloudlets as the input to give the new time quantum for Q_j where $Q_j < Q_i$ in terms of priority.

$$Q_d(j) = S_q*(Q_d(i) + N) \quad (7)$$

Where N is the number of cloudlets in the queue Q_j. Q_j is a lower priority queue and will execute after the higher priority queue Q_i.

C. Quality of Service

The Service Level Agreement that takes place between the Cloud Service Provider and the client is a lawful agreement which specifies the service quality that has to be maintained by the CSP like the maximum delay allowed and pricing model for the services offered. In the proposed work, the Quality of Service can be increased by allowing the user to specify if the request is to be made time efficient or cost efficient. Time efficient requests are those which have to be processed earlier and should have minimum possible delay. Cost efficient requests are those whose processing should fall under certain cost and such request can be delayed to a certain extent but definitely not more than the maximum allowed delay specified under the SLA.

1) Global view: The processing of the cloudlets by the proposed algorithm takes in the complete view or global view of the highest priority queue. The global view keeps a track of the majority of requests in the complete service workload. The algorithm makes the processing time or cost efficient based upon the rate at which each type of request is received.

Let y be the type of request, it has value 1 if the request is time efficient and 0 if the request is cost efficient. P_y is the rate of request of type y in the service workload. Service Workload consists of total number of cloudlets received at the starting, before the training of the model began. Since it takes in consideration the entire service workload, the value provided helps us to have the complete global view of the workload.

$$P_y = \text{count}\{ C(id,f)== y\}/\text{count}\{\text{service workload}\} \quad (8)$$

Where count() is a function that returns the number of cloudlets satisfying the input condition. Service workload is the total number of cloudlets that were received before the training of the model began. C(id,f) is a function that returns the value stored in the flag of the cloudlet with id as ‘id’ and y represents the type of request made, cost or time efficient.

2) Local View: The highest priority queue is made up of entire service workload in the starting. The Highest priority queue is cut into batches of size NVM. This is necessary because the model has a maximum of NVM number of virtual machines available at a time for the processing of the cloudlets.

$$\text{nbatches} = N/NVM \quad (9)$$

if(N%NVM >0)

nbatches= nbatches+1

Each batch is sorted on the basis of the local view.. For each batch F_y is the rate of the request of type y in the particular batch ‘b’.

$$F_y(b) = \text{count}\{ C(id,b,f)== y\}/ \text{count}\{ C(id, b)\} \quad (10)$$

Where, $F_y(b)$ is the rate of request of type y in a particular batch ‘b’. Count() as earlier return the number of cloudlet that fulfils the input parameters. C(id,b,f) returns the flag value of the cloudlet with id as ‘id’ and in the bth batch.

3) Sorting: The Cloudlets in a particular batch are assigned Virtual Machines in a particular iteration in the order of occurrence in the batch itself. For this we calculate a parameter O_y for each type of request y for a particular batch b.

$$O_y(b) = F_y(b) * P_y \quad (11)$$

If(O_y(b) y€0 > O_y(b) y€1) then

Append the request type 1 cloudlet to the request type 0 cloudlet

else

Append the request type 0 cloudlet to the request type 1 cloudlet

Where, cloudlets in a particular request type are sorted individually on the basis of size in descending order. The cloudlet requests type which hold majority in a particular batch are kept first in the batch. This ensures that the cloudlet of that request type is processed earlier than the rest.

D. A* Algorithm

The Artificial Intelligence based A* algorithm provides us with optimal solutions. It is a crossover of the Best First

Solution and the Branch and Bound Method. The Best First Search which utilizes the Heuristic function drives the search towards the goal which is to completely execute a cloudlet. The Branch and Bound method helps to find the partial solutions which are optimal in nature.

After every queue is cut into batches, these batches enter the model one by one. When one batch completes a single tour only then the next batch enters. Each batch in order to complete the tour has to go through NVM number of iterations. Where, NVM are the number of Virtual Machines. In each iteration the cloudlets of the batch are assigned to the Virtual Machine using the A* algorithm. To assign the cloudlets the A* algorithm makes use of Ant Colony Optimization method to find the partial solution and Heuristic function to drive the cloudlet to its completion. Using both of these together the A* algorithm creates an optimal path for the cloudlet. Eventually the cloudlets with similar cloudlet length will follow a same optimized path found by the A* algorithm.

1) Heuristic Function: A heuristic function is a function that takes a particular state as input and gives us a measure on the basis of which we can compare and make a move to the future state. The heuristic function makes the algorithm head towards the goal and exploits the domain knowledge in order to make a move.

In the proposed work, we propose a new heuristic function that drives the search towards the goal that is to completely execute the cloudlet. The heuristic function is a probabilistic value within the universe of [0,1].

$$P_{c,K(i),K+1(j)}(h) = [n_{c,k(i),K+1(j)} / \sum_{j\in\{allowed\ VM\}} n_{c,k(i),K+1(j)}]^{\alpha} \quad (12)$$

$$P_{c,K(i),K+1(j)}(h) = 0 \text{ if } j \in \text{Tabu VMs} \quad (13)$$

Where k is the iteration in which the cloudlet previously was and K+1 is the current iteration. K(i) is the ith VM on which the cloudlet is in the previous iteration and K+1(j) is the is the jth VM on which the cloudlet can be executed in the current iteration and C represents the cloudlet. α is a parameter that controls the effect of the heuristic function. It lies in the universe of [0,1].

Allowed VM= {VM_0, VM_1, VM_2.....VM_{NVM-1}}- Tabu VM's. Tabu VMs are the VMs or cities that a particular cloudlet cannot visit further. Tabu VM's= {C_{rvv} U (VM_s== false)}

Where C_{rvv} are the VM_s that have been visited by the cloudlet in the current tour of the model denoted by 'rvv' which stands for recent VMs visited in the current tour. While VMs represent the status of the VM, it shows if the VM is available to be assigned or not. If the VM has already been assigned to some other cloudlet in the batch, its status becomes false and it cannot be assigned to any other cloudlet in this iteration.

Visibility $n_{K(i),k+1(j)}$ is calculated within the function. It is the inverse of the expected execution time of the cloudlet $d_{K(i),k+1(j)}$.

$$n_{k(i),K+1(j)} = 1/ d_{K(i),K+1(j)} \quad (14)$$

$$d_{c,K(i),K+1(j)} = C(id, l)/ VM(j, m) * VM(j, pe) \quad (15)$$

2) Ant Colony Optimization: The ant colony optimization is a soft computing technique that derives its inspiration from the ants. Ants travel from their nests in search of food when they find the food source they travel back to the nest depositing a pheromone on the way so that other ants can follow the path and reach the food source. The probability that a particular VM in the current iteration will be chosen depends upon the pheromone trail that exists between the VM in the previous iteration to the allowed VM in the current iteration.

$$P_{c,K(i),K+1(j)}(g) = [\tau_{c,k(i),K+1(j)} / \sum_{j\in\{Allowed\ VM\}} \tau_{c,k(i),K+1(j)}]^{\beta} \quad (16)$$

$$P_{c,K(i),K+1(j)}(g) = 0 \text{ if } j \in \text{Tabu VMs} \quad (17)$$

Where K, and C have same meaning as in eq. 12. β is a parameter that controls the effect of the ant colony function. It lies in the universe of [0,1].

a) Pheromone Updating: The pheromone deposited by the ants evaporates at certain rate. This is necessary so that the lesser paths get worn out and eventually the optimal path is found. As the training proceeds the model soon learns the optimal paths that should be followed by the cloudlets of a certain size.

Initially small amount of pheromone τ_0 is deposited on the every link in the graph. The pheromone is updated at two points. Firstly, when the cloudlet finishes before completion of all iteration in the tour. Secondly, when the cloudlet completes the last iteration and is ready to exit the tour. All the connections between the VMs returned by the C(id,rvv) are updated by making a tour backwards till the 0th iteration is reached.

While(K>=1) then

$$\Delta\tau_{k-1(i),K(j)} = Q/ C(id, e) \text{ if } j, i \in C(id, rvv) \quad (18)$$

$$\tau_{K-1(i).K(j)} = \tau_{K-1(i),k(j)} + \Delta\tau_{K-1(i),K(j)} \quad (19)$$

K= k-1

end while

Where K is the current iteration and j is the VM which was visited by the cloudlet in the current iteration and i is the VM visited by the cloudlet in k-1 th iteration. Q is an adaptive parameter. Δτ is the amount of pheromone deposited. After every tour ends or a cloudlet ends the pheromone for the entire graph is updated by the equation 19 below. Fig.2. below shows how the updating takes place.

K= NVM-1

While(K>=1) then

$$\tau_{K-1(i).K(j)} = (1- \mu) * \tau_{K-1(i),K(j)} \quad (20)$$

k= k-1

end while.

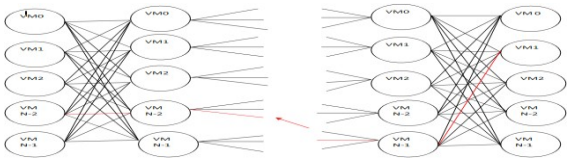


Fig..2. Updating the pheromone trail by backward movement

3) Selection: Once the value of the Heuristic and Ant colony is found for the cloudlet c to all the allowed VMs. Through eq. 20 the value of P c, K(i),K+1(j) (f) is found.

$$P_{c,K(i),K+1(j)}(f)= P_{c,K(i),K+1(j)}(h)+ P_{c,K(i),K+1(j}(g) \quad (21)$$

Where K, and C have same meaning as in eq. 12. After this the VM with maximum P(f) value for the cloudlet c, is selected to be run upon it as per the eq. 21. After this the status of the selected VM is changed to false so that no other cloudlet in the batch can choose it and also the VM id is added to the list of recently visited VMs and visited VMs of the cloudlet c.

$$\text{selectedvm}= \prod_j\{ \max \{ P_{c,K(i),K+1(j)}(f)\}\} \quad (22)$$

$$\text{if VM(id==selectedvm) then VM(id,s)= false} \quad (23)$$

$$C_{rvv}.\text{add(selectedvm)} \quad (24)$$

$$C_{vv}.\text{add(selectedvm)} \quad (25)$$

Where the add() is a java array list function that adds the selected VM to the list of the recently visited and visited VM.

4) Running: Once all the cloudlets are assigned a Virtual Machine in a particular iteration. The cloudlets are run upon their selected Virtual Machines for a maximum time period equivalent to the time quantum.

$$\text{Selectedvm= C(id, rvv).lastof()} \quad (26)$$

$$\text{C(id,l)= C(id,l)- VM((id==selecetdvm), m)*C(id,t)} \quad (27)$$

$$\text{C(id, e) = C(id,e) + C(id,t)} \quad (28)$$

Where C(id, l) return the cloudlet length and C(id,t) return the time period for which the cloudlet can run upon the selected VM j. VM((id==j), m) returns the processing capacity in Million Instructions per second of the VM j. After all the cloudlets run upon their designated VMs the iteration ends and if the cloudlet length is left it moves to the next iteration where again using the A* algorithm again the VM are selected.

If all the cloudlets are executed in a particular iteration and the tour isn't complete the next batch begins its tour from the starting.

If some cloudlets are executed but the tour is still ongoing for the rest of the cloudlets. The VMs to which cloudlet are not scheduled in any iteration, go to sleep mode to save power.

If all the cloudlets are not executed and the tour is complete i.e. NVM number of iterations have been performed the remaining cloudlets move into a lower priority queue.

When all the batches complete the tour and the lower priority queue isn't empty. The time quantum for the lower priority queue is calculated using the 3.2.1.4 Dynamic module and cut into batches and then enters the model.

If the lower priority queue becomes empty the cloudlets have been executed completely and hence the algorithm terminates.

After completion of iteration in every tour, the cloudlets that lie in the lower priority queue have their waiting time increased by the value of the time quantum.

5) Working of the Algorithm:

1. The service workload is added to the first queue and the time quantum for the queue is found using the Inference Module.
2. Then the Queue is cut into batches, sorted and scheduled for their turn into the model as per Sorting.
3. The batches enter the model one by one and each batch makes a single tour in which it passes through NMV number of iterations. In iteration all the cloudlets are assigned their respective VM using the Eq. 22 in Selection.
4. These cloudlets are then executed upon the selected JVM as per the Eq. 27 and Eq. 28 in Running.
5. The cloudlet if completely executed updates the pheromone on the basis of the Pheromone Updating if not it increments the iteration goes back to the step 3.
6. If the number of iteration is equal to NVM-1 then update the pheromone trail and move the cloudlet to the lower priority queue.
7. Begin the tour again from step 3 for the next batch and execute the cloudlets.
8. If all the batches are executed then move to the lower priority and calculate time quantum using the Dynamic Module and go to step 2.
9. If there exists no lower priority queue the cloudlets are completely executed, stop.

IV. Experimentation and Result

The performance of the proposed Machine Learning Model is compared and analyzed in terms of completion time and Quality of Service. The ML model is simulated in java environment. The performance is measured by setting up different simulation environments with varying number of cloudlets and the Virtual Machines.

A. Parameters Setup

Q which is an adaptive parameter has been set to 100. $\alpha=\beta= 1$ both the heuristic and the Ant colony have equal effects in the VM selection. $\tau_0 =1$ is the initial pheromone deposited upon the path.

B. *Completion time*

The cloudlets are compared on the basis of the completion time. The total time taken by the cloudlets is the sum of execution time and waiting time of the cloudlets. Difference in total time is the difference between the total time taken by current and previous service workloads.

Cloudlets	waiting time average	execution time average	total time average	difference in total time average
501	68.76	10.57	79.33	0
1001	146.44	10.80	157.24	77.91
1501	222.00	10.72	232.72	75.47
2001	296.07	10.76	306.83	74.11
2501	376.14	10.91	387.06	80.22
3001	457.84	10.96	468.80	81.74

TABLE I. COMPLETION TIME OF CLOUDLETS

The results reveal that the average execution time of the cloudlets (time taken to execute upon the Virtual Machine) increases minutely with increase in service workload. Also, difference in total time of the service workloads remains approximately similar every time. This shows that the model is able to train itself effectively with the increase in service workload.

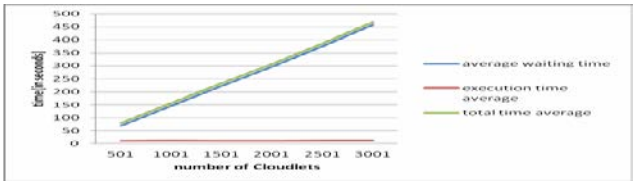


Fig.3. The execution time of cloudlets as per the proposed algorithm

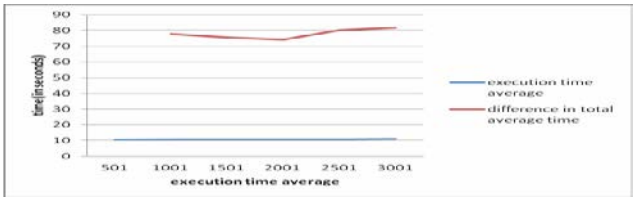


Fig.4. Execution time of cloudlets as per proposed algorithm

C. *Quality of Service*

The Quality of the service for the cloudlets is maintained by allowing the user to specify whether it wants the request to be time efficient or cost efficient. As shown in Fig.5. the algorithm in starting fluctuates greatly in term of the execution time of both cost and time efficient cloudlets. But eventually converges and the execution time decreases drastically for the cloudlets. While the execution time for the time efficient cloudlet decreases much more in respect to the execution time of the cost efficient cloudlets.

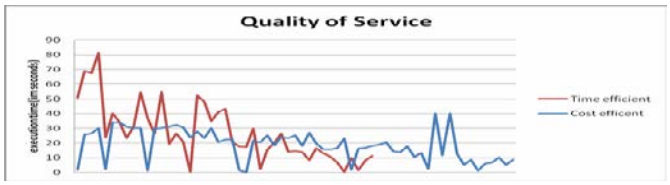


Fig.5. Execution time of cloudlets in respect to Quality of Service.

V. CONCLUSIONS

This work highlights a Machine learning technique which makes use of various concepts like Multilevel Feedback Queue, travelling Salesman, Ant Colony Optimization, Heuristic, and A* algorithm. The model eventually is trained and converges to follow an optimal path depending upon the instruction length of the cloudlets. The future scope of this work involves creation of tradeoff between the Heuristic and Ant Colony values in the A* algorithm.

REFERENCES

[1] S. Parsa, R. Entezari-Maleki. “RASA: a new grid cloudlet scheduling algorithm” in World Appl. Sci. J. 7, 152–160 (Special Issue of Computer & IT).

[2] Chen Liu, Qiu, Cai & Huang.”Scheduling Parallel Jobs using Migration & Consolidation in the Cloud” in Hindwai Publications of Mathematical Problems in Engineering, July 2012.

[3] H. Goudarzi, M. Ghasemazar, and M. Pedram, “Sla-based optimization of power and migration cost in cloud computing,” in Cluster, Cloud and Grid Computing (CCGrid), 2012 12th IEEE/ACM International Symposium on. IEEE, 2012, pp. 172–179.

[4] Bajoria V., Katal A. “CAMQU: A Cloudlet Allocation Strategy Using Multilevel Queue and User Factor”. In: Bhattacharyya P., Sastry H., Marriboyina V., Sharma R. (eds) Smart and Innovative Trends in Next Generation Computing Technologies. NGCT 2017.

[5] K. Nishant, P. Sharma, V. Krishna, C. Gupta, K. P. Singh, R. Rastogi et al., “Load balancing of nodes in cloud using ant colony optimization,” in Computer Modelling and Simulation(UKSim),2012 UKSim 14th International Conference on. IEEE, 2012, pp. 3–8.

[6] R. Shojaee, H. R. Faragardi, S. Alaee, and N. Yazdani, “A new cat swarm optimization based algorithm for reliability-oriented task allocation in distributed systems,”in Telecommunications(IST), 2012 Sixth International Symposium on. IEEE, 2012, pp. 861–866.

[7] S.G.Domanal and G.R.M.Reddy,“ Load balancing in cloud computing using modified throttled algorithm,” in Cloud Computing in Emerging Markets (CCEM), 2013 IEEE International Conference on. IEEE, 2013, pp. 1–5.

[8] S. Domanal and G. Reddy, “Optimal load balancingin cloud computing by efficient utilization of virtual machines,” in 2014 Sixth International Conference on Communication Systems and Networks (COMSNETS). IEEE, 2014, pp. 1–4.

[9] G.Gonc¸alves,P.Endo,M.Santos,D.Sadok,J.Kelner,B.Melander, and J.-E. Mangs, “Cloudml: An integrated language for resource, service and

request description for d-clouds," in Cloud Computing Technology and Science (CloudCom), 2011 IEEE Third International Conference on. IEEE, 2011, pp. 399–406.

[10] W. Zhou, S. Yang, J. Fang, X. Niu, and H. Song, "Vmctune: A load balancing scheme for virtual machine cluster using dynamic resource allocation," in 2010 Ninth International Conference.

Analysis of cardiovascular diseases using artificial neural network

Jyotismita Talukdar, Bhupesh Kumar Dewangan
University of Petroleum and Energy studies, Dehradun, India
jtalukdar@ddn.upes.ac.in, b.dewangan@ddn.upes.ac.in

Abstract: In this paper, a study has been made on the possibility and accuracy of early prediction of several Heart Disease using Artificial Neural Network. (ANN).The study has been made in both noise free and noisy environment. The data collected for this analysis are from five Hospitals. Around 1500 heart patient's data has been collected and studied. The data is analysed and the results have been compared with the Doctor's diagnosis. It is found that, in noise free environment, the accuracy varies from 74% to 92%.and in noisy environment (2dB),the results of accuracy varies from 62% to 82%. In the present study, four basic attributes considered are Blood Pressure (BP), Fasting Blood Sugar (FBS), Thalach (THAL) and Cholesterol (CHOL.).It has been found that highest accuracy(93%), has been achieved in case of PPI(Post-Permanent-Pacemaker Implementation), around 79% in case of CAD(Coronary Artery disease),87% in DCM(Dilated Cardiomyopathy), 89% in case of RHD&MS(Rheumatic heart disease with Mitral Stenosis), 75 % in case of RBBB +LAFB (Right Bundle Branch Block + Left Anterior Fascicular Block),72% for CHB(Complete Heart Block) etc. The lowest accuracy has been obtained in case of ICMP(Ischemic Cardiomyopathy),about 38% and AF(Atrial Fibrillation) , about 60 to 62%.

Keyword: Coronary Heart Disease, Cardiovascular Disease, Thalach, Cholesterol, (Sick Sinus Syndrome (SSS), Chronic Stable Angina (CSA).

I. INTRODUCTION

One of major challenge facing the healthcare organization today is the gift of value administrations at sensible expenses. The nature of administration of administration infers diagnosing and directing the patients accurately. A lion's share of territories of wellbeing administrations, for example, conjecture of heart assault, effectiveness of surgeries, therapeutic tests, prescription and effectiveness of medical handlings can be estimated by the application of ANN. Artificial Neural Network (ANN) has extensive application to biomedical systems. Neural networks acquire knowledge by themselves. Due to this, they are able to identify the diseases that are more susceptible. It requires only set of examples to represent all the variations of the diseases possible.

Experimentally, the methodology of neural systems are utilized to demonstrate the human cardiovascular framework. By building a sham of the cardiovascular arrangement of an individual, afterward contrasting, and the constant physiological estimations, for example, heart rate, Blood weight, Blood sugar, Cholesterol and so on taking from the patients, we can make an early prediction of the disease . If the model is found to be adjusted to an individual, at that point it turns into a model of that person. Determination of coronary illness is a troublesome and monotonous errand in medicinal field. When all is said in done, every one of the specialists are anticipating coronary illness by learning and experience. The analysis of coronary illness is a multi-layered issue which once in a while may prompt false capricious impacts. Some of major issues related to correct diagnosis of heart disease are:

- Less precise outcomes,
- Less experience,
- Time subordinate execution ,
- Knowledge up degree,
- Complex and multiplexed side effects and so forth.

Hian Chye and Gerald Tan[12] suggested that if the clinical choice help and the PC construct displaying with respect to understanding records work in a coordinated way then the it could lessen the restorative blunders, upgrade persistent security, diminish undesirable practice variety. This would surely enhance the patient result. In perspective of this the ANN can possibly produce a learning rich condition, which can have huge impact on expanding the nature of clinical choices. An Artificial neural system is a multilayer organize comprised of information layer neurons, shrouded neurons and yield neurons.[1][2]. They are considered as proficient methods for expectation, streamlining and acknowledgment that are troublesome for customary PCs or human beings.[3]. They are utilized for non-parametric forecast that gain from the environment, hold the learning and later utilize it accordingly. Essentially, the Artificial Neural Networks are constituted by the arrangement of interconnected gatherings of fake neurons and data handling units which are grouped following a connectionist approach**.[4]**. It demonstrates non-linear processing and broadly used in performing pattern matching, prediction and recognition etc. based on a method that uses continuously updated connectionist weights during learning and training. ANNs are efficient means to relate input data to the expected class decisions. They are basically adaptive networks by origin and keep changing their own internal structures and information which flows along the system during their training stages [5]. The important advantages of ANNs are:

1. They are more robust , even in noisy environment, because of weights.
2. ANN enhances its execution by realizing which proceeds notwithstanding amid the preparation stage too.

3. ANN can be parallelized for better execution.
4. There is low mistake rate and once the fitting preparing has been performed

An Artificial neural system have a parallel , disseminated data preparing structure comprising of different number of handling components, called Nodes and they are interconnected. . Each handling component has a solitary yield association that branches into numerous different associations conveying a similar preparing component yield flag. A basic model of a run of the mill organic neuron is appeared in Fig.(1.0) below.

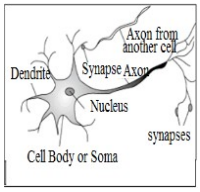


Fig (1): Symbolic representation of biological neuron.

The key data handling unit of the ANN is the McCulloch-Pitts Neuron (1943) [6], like the natural case. Figure (2.0) demonstrates the model of a McCulloch-Pitts neuron utilized for planning ANNs.

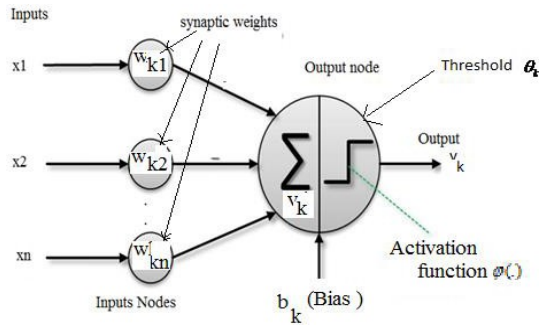


Fig (2): McCulloch-Pitts neuron Model

II. ELEMENTS OF NEURAL NETWORK

The three fundamental components of ANN or NN – show are:

1. An arrangement of neural connections or interfacing joins, every one of which is described by weight or quality of its own. For instance, a flag xj at the contribution of neurotransmitter j associated with v neuron , k is increased by the synaptic weight w kj, alludes to the neuron being referred to and the second subscript alludes to the info end of the neurotransmitter [7].
2. An snake to summing all the info signals, which are weighted by the individual neurotransmitters of the neuron.
3. To limit the amplitude of the output of a neuron, an activation function is used.

III. CLASSIFICATION OF ARTIFICIAL NEURAL NETWORK

The ANN can be ordered in two fundamental gatherings as indicated by the manner in which they learn. They are:

1. Supervised learning.
2. Unsupervised learning.

In administered taking in, the systems register a reaction to each information and contrast and the objective esteem. On the off chance that the information reaction varies from the objective esteem then the weights of the system are adjusted by a learning guideline for instance: Single-Layer Perception (SLP) Multi-Layer Perception (MLP) and so on. The graphical perspective of such learning is appeared in Fig. (3) beneath.

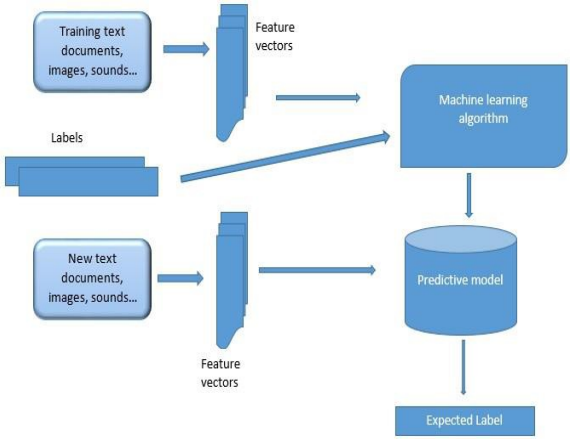


Fig.(3) : Supervised Learning Model

In the event of Unsupervised taking in the systems learn by distinguishing extraordinary highlights in the issues they are presented to , i.e Self-Organizing Features Maps(SOM). A common Block portrayal of unsupervised learning has been appeared in Fig (4) underneath

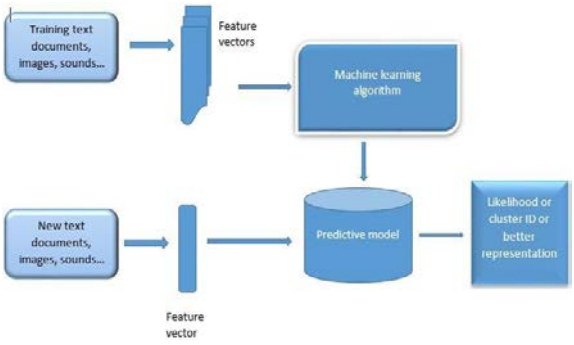


Fig 4: Block representation of Unsupervised Learning.

The ANNs are designed based on the Cerebral Cortex architecture of human brain. There are two types of ANN. They are-

- Feed-forward ANN,
- Recurrent or Feed-backward ANN.

The Feed-Forward ANN process data applies forward pass and then produces the outcome. It has varied number of layers and having single-layered or multi-layered forms. The number of layer is related to layers of artificial neurons in the ANN. In the Feed-Back ANN (FBANN) or Recurrent Neural Network (RNN) the solutions are unknown. In such a network, the data propagation takes place in backward directio[5]. A recurrent ANN can utilize their internal memory to process any sequence of inputs. Some of the most commonly used features of Recurrent ANN are :- Hopfield, Jordan and bi-directional etc.. A typical block representation of Recurrent Neural Network is shown in Fig (5) below.

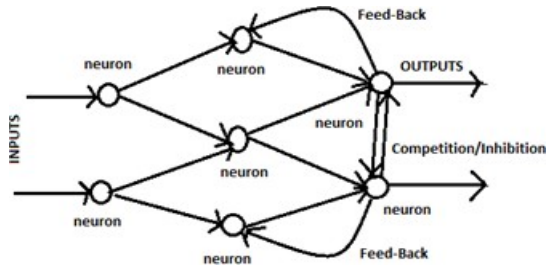


Fig (5): Block representation of Recurrent Neural Network

IV. TRAINING OF ARTIFICIAL NEURAL NETWORK

The preparation of the ANN is finished after two Passes, (I) A Forward Pass, and (ii) a regressive calculation with blunder assurance and interfacing weight refreshing in the middle. It is normal that the preparation must be attempted to quicken the speed of preparing and the rate of meeting of the Mean Square Error (MSE) to the coveted esteem [7]. The consecutive the means are:

1. Initialization of weight matrix, and
2. Presenting the Training Samples.

In the present study, we have primarily used the Matlab to define a MLP.

To fit with the property of ANN having m-layers of functional units, i.e (m-1) hidden layers, the Matlab function used is defined as –

ffnet=newff(input_data,desired_output $\{k_1,k_2,.....,k_{m-1}\},\{f_1,f_2,.....,f_m\}$, '<Training algorithm>').

The parameters involved with newff( ) are as follows

1. Input_Data, [X]: A matrix with input data through its columns in the training set which decides the number of input units for newff( ).
2. Desired_Output, [T]: A matrix with the correct answers through its columns in the training set. It decides the number of output units for newff ( ).

The commonly used neural network or artificial neural network supervised training algorithm is the Back-Propagation Algorithm. The training of a NN/ANN by Back-Propagation algorithm includes the following three stages:

(i). Feed forward algorithm,
(ii). Analysis and back propagation of the associated error.
(iii). Weight balancing.

In the current analysis of CVD/CHD and their early prediction, we have used the Back-Propagation Algorithm. While training the MLP we have used two training functions:

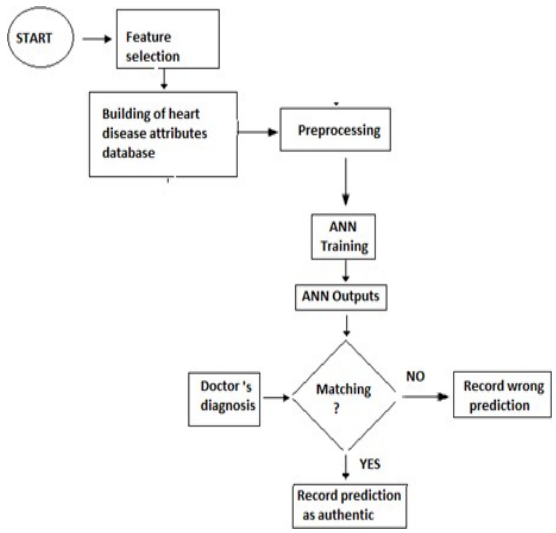
(i) adapt () and (**ii) train ().**

V. SPECIFICATION OF ANN

In the current study the back propagation algorithm using feed forward ANN is analyzed for the classification and recognition of CVD. In this methodology, we have taken the ANN layer where input is of 12 neurons and is tried for 200 to 1000 repetitions. The values achieved are the mid values of at least 10 to 15 tests for the times measured. Here, the Artificial NN have considered two hidden layers and its primary terms are provided in table IV. The ANN is organized in such a way that it can take care of size and SNR differences. The feature sizes varies from (1050 x 4) for training, (225 x 4) for testing and (225 x 4) for validation check. In all the testing stages some

noise have been added with SNR varying from 0 to 3dB. While testing the ANN, we have used error back propagation algorithm coupled to gradient descent with Levenberg-Marquardt(LM) [8][9]10][12] optimization ,which is fast and suitable for supervised training. It is found that the LM optimized back propagation gives efficient learning despite variations in patterns(during training, testing and validation), though it consumes a little bit of more memory.

The essential steps considered to apply the concepts of Artificial Neural Networks for the early detection of occurrence of Cardiovascular diseases with satisfactory confidence is shown in Fig.(6)below:



Fig(6): Flow diagram of ANN algorithm

While designing the algorithm it is assumed an initial momentum value (around 0.001) with a decreasing factor of 0.1 and increasing step of 10. The minimum performance gradient is taken as 10^{-7}. The training methods of ANN considered for the present work are as follows:

1. Back-propagation(BP) with Gradient Descent(GD): The BP algorithm is used to learn the weights of a MLP which is a static in nature. Through the gradient descent it shrinks the squared summation inaccuracy between the network's final values and the given target values [7].
2. BP with Momentum (M): By changing the weight the convergence is improved. This is achieved by the observation of BP based on the modification on the momentum [12].
3. BP with GD and M and varying learning rate (LR): The LR parameter determines the speed of the BP to reach the convergence. When the learning rate is higher, the size of each step will be bigger and in turn the convergence will speed up. [11] [13].

VI. DATA COLLECTION

The data used in the present study is collected from five medical healthcare organizations, including Govt. Hospitals and Private Nursing Home , all located at and around the greater Guwahati(this has already been mentioned in chapter-III), which is the prime healthcare centre of the entire north-east region. There are , total 76 attributes in the medical database , but in the present study of Cardiovascular Disease(CVD)/Coronary Heart Disease(CHD) using ANN we are taking only 08 attributes. They are shown in Table(I) below :

TABLE I
ATTRIBUTES OF THE MEDICAL DATABASE

Serial No.	Attribute Name
1	Age
2	Sex
3	Chest Pain(CP)
4	Trestbps(BP)
5	Cholesterol(CHOL)
6	FBS
7	Thalach(THAL)
8	Heredity

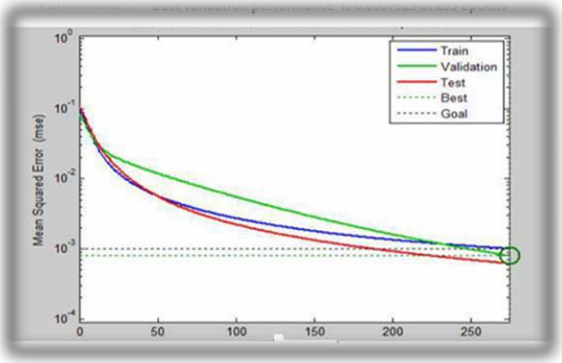
VII. EXPERIMENTAL DETAILS AND RESULTS

In the present study of early prediction & detection of Cardiovascular Disease (CVD)and Coronary Heart Disease(CHD) using ANN, a total of 1500 heart patients have been studied. The records of the patients have been collected from the Govt. Hospitals and Private Nursing Homes, as mentioned in chapter-III. The prediction using ANN have been made using four primary heart attributes, namely: - Blood Pressure(BP), Fasting Blood Sugar(FBS) Thalach(THAL.) and Cholesterol(CHOL.). With respect to a particular heart disease, no distinction has been made with respect to sex. Further, in the present analysis age and family history are also not considered as cues causing heart diseases. In the present study of early prediction & detection of Cardiovascular Disease (CVD)and Coronary Heart Disease(CHD) using ANN, a total of 1500 heart patients have been studied. The records of the patients have been collected from the Govt. Hospitals and Private Nursing Homes. The prediction using ANN have been made using four primary heart attributes, namely: - **Blood**

Pressure(BP), Fasting Blood Sugar(FBS) Thalach(THAL.) and Cholesterol(CHOL.).

In the present study of heart disease using ANN, out of 1500 heart patient's data, 70% of data has been used for training the ANN, 15% data used for validation test and rest 15% of data used for testing the ANN's performance.

70**% (1050**) of the heart disease data, with four major attributes i.e BP.FBS, THAL, and CHOL. have been used for training the ANN. **15% (225**) data(BP,FBS,THAL.,CHOL) have been used for validation test and rest **15% (225)** data of {BP,FBS,THAL. and CHOL } have used for testing the ANN. A typical Snapshot of regression plot has been shown in Fig.(7) below.



Fig(7): Representation of regression plots

The confusion matrix for success rate achieved during training and testing have been shown in Fig.(8), below.

Fig (8): Confusion matrix with 60% success in training set

It is thus seen from the graphical outputs of the three phases of ANN, i.e Training, Validation and Testing, the following observations have been made:

1. The best performance is achieved at error 0.001, which is as our initial assumption (goal).
2. During Training, nearly 60%to 80% success has been achieved, as presented in the confusion matrix.
3. During validation, about 66.7% success has been observed.

Further, the exhibition of the ANN utilizing Back-proliferation calculation is likewise assessed by figuring the rates of Sensitivity(SE), Specificity(SP) and accuracy(AC) utilizing the accompanying computational number juggling [15]. These parameters have been registered utilizing the Confusion grid.

TABLE II
STRUCTURE OF CONFUSION MATRIX

		Actual Value		
		p	n	Total
Prediction	P'	TP	FP	P'
	n'	FN	TN	N'
	Total	P	N	

$$\mathbf{SE=}\frac{\boldsymbol{True\ Positive}}{(\boldsymbol{True\ Positive}+\boldsymbol{False\ Negative})}\mathbf{*\ 100}$$

$$\mathbf{SP=}\frac{\boldsymbol{True\ Negative}}{(\boldsymbol{True\ Negative}+\boldsymbol{False\ Negative})}\mathbf{*\ 100}$$

$$\mathbf{AC=}\frac{(\boldsymbol{True\ Positive}+\boldsymbol{True\ Negative})}{(\boldsymbol{True\ Negative}+\boldsymbol{True\ Positive}+\boldsymbol{False\ Negative}+\boldsymbol{False\ Positive})}\mathbf{*\ 100}$$

Out of 1500 patient data on various health attributes, 70% have been used to train the algorithm, and the remaining 30% have been used for testing the performances. Following the steps, as mentioned in TABLE III the results obtained on SE, SP and AC have been shown.

TABLE III
RESULTS OF ANALYSIS

Algorithm used	Sensitivity	Specificity	Accuracy
MLPANN with Back Propagation Algorithm	82%	87%	81%

It is thus seen from the above results that Heart Disease Analysis system using Artificial NN with the back propagation gives better accuracy, sensitivity and specificity when compared to other methodologies, like Statistical and Data mining.

References:

[1] A.K.Nanila et. Al, "Fault diagnosis of mixed- signal analog circuit using Artificial Neural Network." Int. J. of Intelligent Systems and Applications. 07,pp.11-17,2015.

[2] S.Bhuvaneswari, J Sabarathinam; " Defect Analysis using Artificial Neural Network." Int. Journal of Intelligent Systems and Applications, 05, pp.33-38 , 2013.

[3] L.Fausett : "Fundamentals of Networks Architectures, Algorithms and Applications. " Pearson Education, 1993.

[4] S.Mitra and Y.Hayashi : " Neuro-Fuzzy rulegeneration: Survey in Softcomputing Framework." IEEE Transactions on Neural Networks, Vol.2,No. 3,pp.748-768,2000.

[5] Kausik Roy,Dipankar Das and M.Ganjer Ali : " Development of the speech recognition system using ANN." 5th ICCIT 2002,East-WestUniversity, Dhake,Bangladesh, 27-28,Dec. 2002.

[6] McCulloch W.S and Pitts W : " A logical calculus of the ideas imminent in nervous activity." Bull Math. Biophy.5: 115-133,1943.

[7] S.Haykin "Neural-Networks: A comprehensive Foundation." New Delhi, Pearson Education, 2nd. Edition, 2003.

[8] B.M Wilamowski : " Neural Network Architectures and learning algorithms." IEEE Industrial Electronics Magazine." Vol.3, no.4, pp.56-63,2009.

[9] K. Levenberg : "A method for the solution of certain problems in Least Squares." Quaterly of applied mathematics, vol.5,pp.164-168,1944.

[10] D.Marquardt : " An Algorithm for Least- Squares Estimation of Non-linear parameters." SIAM Journal . on Applied Mathematics, vol.11 no.2, pp.431-441 , 1963.

[11] J.Principe, N.Euliano and W.Lefebvre : " Neural and Adaptive system- Fundamentals through simulations." Wily,2000.

[12] Hian Chye Koh and Gerald Tan : " Data mining applications in Healthcare." J. of Healthcare information management. vol.19, issue-02, pp.64-72,2005.

[13] A.T.Sayad and P.P. Halkarnikar : " Diagnosis of heart disease using Neural Network Approach. " Int. J. of Advances in Science and Engineering and Technology. Vol.02,Issue-03,2014.

[14] R.H.Nielsen : " Theory of the True Back Propagation Neural Network." Proc.IEEE IJCNN,pp. 1593-1605,IEEE press,NewYork, 1989.

[15] A Literature Review of Cardiovascular Disease Management Programs in Managed Care Populations,SHETA ARA, PharmD,

Automated Irrigation System-IoT Based Approach

Dweepayan Mishra[1],Arzeena Khan[2] Rajeev Tiwari[3], Shuchi Upadhay[4]
UPES[1,2,3], UCALS[4]
Khan.arzeena@gmail.com

Abstract: **Agriculture is a major source of earning of Indians and agriculture has made a big impact on India's economy. The development of crops for a better yield and quality deliver is exceptionally required. So suitable conditions and suitable moisture in beds of crop can play a major role for production.. Mostly irrigation is done by tradition methods of stream flows from one end to other. Such supply may leave varied moisture levels in filed. The administration of the water system can be enhanced utilizing programmed watering framework This paper proposes a programmed water system with framework for the terrains which will reduce manual labour and optimizing water usage increasing productivity of crops. For formulating the setup, Arduino kit is used with moisture sensor with Wi-Fi module. Our experimental setup is connected with cloud framework and data is acquisition is done. Then data is analysed by cloud services and appropriate recommendations are given.**

I. INTRODUCTION

India is a horticultural nation, where population is over 1.2 billion, out of which around 70% of the population relies upon horticulture. Agriculture is a major source of earning of Indians and agriculture also has made a big impact on India's economy. Agriculturists have an extensive variety of assorted variety to choose reasonable products of the soil crops. Be that as it may, the development of these crops for ideal yield and quality deliver is exceptionally specialized. It can be enhanced by the guide of innovative bolster. The administration of the water system can be enhanced utilizing programmed watering framework This paper proposes a programmed water system with framework for the terrains which will reduce manual labour and optimizing water usage increasing productivity of crops. Presently the computerization is one of the critical parts in the human life which gives comfort as well as lessen burden and helps us to save time We plan to develop a framework that helps the farmer to automatically provide water to the plant according to its need and current water moisture present in the soil. A keen water system is developed with the help of moisture sensors and Arduino chips. In the system, we bury moisture sensor into the soil which would notify the system about amount of water present in the soil. With the help of a program, coded in C language, system will check the amount of water required by a plant, with predefined values in the program. If the moisture level is less than the amount of water needed by the plant, the program automates the flow of water from a submersible pump unless a threshold value is reached. This ensures that crop has been provided optimum amount of water without any manual labour or wastage. It improves efficiency of water usage, reduced cost of irrigation water, intelligent irrigation.

II. EMBEDDED SYSTEMS

Embedded systems are PC systems that is a piece of bigger systems and them play out a portion of the prerequisites of these systems. A few cases of such systems are auto portable control systems; mechanical forms control systems, cell phones, or, on the other hand, little sensor controllers. Embedded systems cover an extensive scope of PC systems from ultra-little PC based gadgets to extensive systems checking and controlling complex procedures. The overpowering number of PC systems has a place with embedded systems: 99% of all registering units have a place with embedded systems today.

A. ARDUINO

Arduino is an electronic platform built on easy to use hardware. It is a open source software. [1] Arduino UNO is one of the most easily available low-cost Arduino board. The Arduino is an embedded system. Various pins on the Arduino are used to read or write values on to the system. Many types of micro controllers are available in the market. Some of them are Parallax Basic Stamp, Netmedia's BX-24, Phidgets, MIT's Handyboard that provide the same functionalities but Arduino holds the following advantage over them:

- Inexpensive
- Cross Platform (Linux, Mac OS, Windows)
- Simple clear programming environment

978-1-5090-6785-5/18/$31.00 © 2018 by IEEE

- Open Source and Extensible software and hardware.

III. LITERATURE REVIEW

There has been quite a few research work going on the agenda of automation of irrigation systems. Various other technologies (different microprocessors, different algorithms) have been used to reach variety of conclusions. Various scientists have worked with programmed water sprinkling or water system framework. They picked distinctive measurements for deciding the soil condition and amount of water. They likewise examined about various wellsprings of energy for the sensors. Plus, the innovation for making system among the sensors and outline of control framework were additionally intensely talked about by the researchers. An article on the mechanized water supply framework for urban local locations appeared that such a framework can be utilized to adequately oversee water asset. [2]

The aim of this system is to modernize farming innovation by using programming segments and construct the necessary parts for the framework. The framework is ceaseless based and focuses the right condition of paddy field. There is one central centre used which to control another centre. The key limit of RF module is to pass the message to the centre point and work the system. [3]

IV. DESIGN OF SYSTEM

A. Soil moisture sensor:

The Soil Moisture Sensor (SMS) is a sensor associated with a water system framework controller that measures soil dampness content in the dynamic root zone before each planned water system occasion and sidesteps the cycle if dampness is over a client characterised set point.

B. Arduino:

Arduino is an open source electronics platform built on easy to use hardware. It comes with its own IDE (Arduino IDE) and its own open source extensible hardware.

C. Use of sensors:

The Soil Moisture Sensor (SMS) is a sensor connected to the irrigation system controller that measures the soil moisture content. The soil moisture sensor reads the value of moisture content in the soil and prints it on the console to view the values. A threshold value is set at the beginning depending upon the plant being watered and region in which it is being watered. The soil moisture sensors reads the value once within a stipulated delay time. Once the moisture is above the user defined threshold value the sensor stops reading the value and the control passes on to the Arduino which switches on the pump to start the watering of the system.

D. Data Acquisitions:

Data Acquisition is the handling of various electrical or electronic contributions from gadgets, for example, sensors, clocks, transfers, and strong state circuits with the end goal of checking, breaking down or potentially controlling frameworks and procedures. Information securing instrument sorts incorporate PC sheets, instruments or frameworks, data loggers or recorders, outline recorders, input modules, yield modules, and I/O modules.
In this research work data is acquired by usage of the Soil Moisture Sensors.

- **Soil Moisture Sensors** works on the principle of Dielectric permittivity. The dielectric permittivity is the amount of electricity that can be passed through the soil. The dielectric permittivity is a function of water content present in the soil. Hence by measuring the dielectric permittivity we could measure the soil moisture content. A fixed (user defined) threshold value is set and data is acquired till it reaches the threshold value. Once it has reached the stipulated value the soil moisture sensor bypasses the reading of the value for one cycle.

E. *Decision regarding threshold value:*

The soil moisture sensor is buried in the soil and water is applied to the soil. At least one inch of standing water is put on the soil.

- The soil along with soil moisture sensor is left outside on the sun for twenty four hours and If it rains within this period the process needs to start over.
- After twenty four hours the value of the soil moisture is read and is set as a threshold value. A 20% reduction can be done on the moisture level to allow a little more time for the water to seep in.

V. WORK FLOW CHART:

IoT based system of irrigation works in cooperation with sensors on Arduino kit. All its functioning is shown in Fig 1. Firstly depending on need of crop a threshold value is set on moisture sensor. Then continuously humidity read by sensor is checked against the threshold values. If humidity value is less then threshold then still

irrigation is continued. When threshold value is reached then pump is switched off automatically by sending signals through Arduino kit.

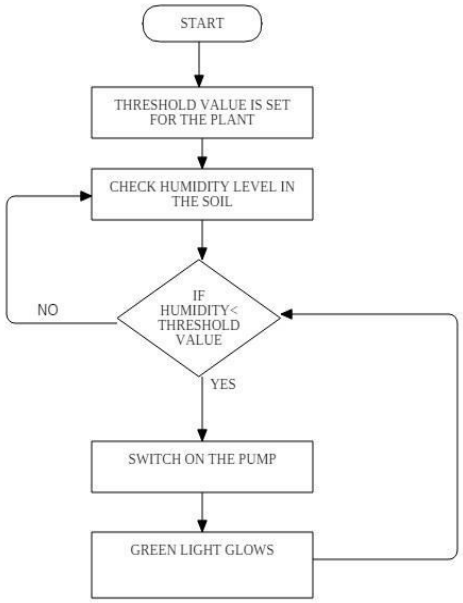


Fig1: Flowchart of process used.

VI. System Implementation:

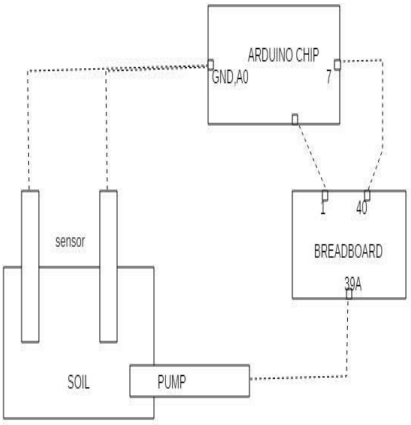


Fig 2: System Implementation

Arduino microprocessor board is connected with a bread board to extend the connections, for connecting the water pump with microprocessor. Soil moisture sensors are connected with Arduino kit to get readings of moisture of soil of farm. Then these gathered values are compared with threshold values of moisture levels and accordingly pump is being operated switched on or off as shown in Figure 2.

VII. DATA ANALYSIS

Data Analysis is a strategy in which information is gathered and composed with the goal that one can get accommodating data from it.

A. Gathering Of Data :

Soil Moisture Sensors works on the principle of Dielectric permittivity. The dielectric permittivity is the amount of electricity that can be passed through the soil. The dielectric permittivity is directly proportional to the amount of water present in the soil. Hence, by measuring the dielectric permittivity we could measure the soil moisture content. Soil Moisture Sensors are buried and are connected to the Arduino chipset at the other end. The soil moisture sensor reads the value of the dielectric permittivity of the soil after a stipulated interval of time. These values are sent to the Arduino chipset and are correspondingly displayed on the system.

B. Analysis Of Data :

A threshold value is set at the beginning of the procedure .The steps before setting up the threshold value are as follows :

- The soil moisture sensor is buried in the soil and the flow of water is opened. At least one inch of water is allowed to stand on the soil.
- The soil is left under the sun for twenty four hours. If it rains in this interval the procedure has to be started from the beginning.
- The moisture value is noted at the end of twenty four hours and with twenty percent deviation from the moisture value, the threshold value is set.

After acquiring the threshold value, the soil moisture sensors are allowed to read the moisture content in every fixed interval of time. The data is gathered and compared o the threshold value. After the comparison, the system has two course of actions:

Case 1: If the moisture content is more than the threshold value.

When the moisture content read from the soil using the soil moisture sensor and it is found out to be more than the threshold value, then after a delay of a fixed time the value is read again and compared. This procedure is continued until Case 2 is encountered.

Case 2: If the moisture content is less than the threshold value.

When the moisture content read from the soil using the soil moisture sensor and it is found out to be less than the threshold value, then the system bypasses one circle of reading the values. Signal is sent to the pump to notify about a state change. (From LOW to HIGH)

Valve for the particular strip in which the moisture content is less than the threshold value is opened and water from the pump is allowed to flow.

CONCLUSION

India is a country with most of agriculture on its land. Irrigation is need of agricultural outputs. Better and optimally farms are irrigated suitable is the crop yield. So this work has designed a smart irrigation system based on IoT with sensor of humidity. It may check the moisture content levels of soil in farm and can generate moisture level data through sensors. Accordingly irrigation based decisions are taken by system automatically to start water pump and to divert the flow of pump motor for irrigation. Designed system can irrigate field with lesser amount of water. Crop can be maintained with its suitable threshold moisture levels for better yields.

REFERENCES

[1] https://www.arduino.cc/en/Guide/Introduction

[2] N.B. Bhawarkar, D.P. Pande, R.S. Sonone, Mohd. Aaquib , P.A. Pandit, and P. D. Patil, “Literature Review for Automated Water Supply with Monitoring the Performance System”, International Journal of Current Engineering and Technology, Vol. 4, No. 5, Oct 2014.

[3] Rane, et al .,"Review Paper Based On Automatic Irrigation System Based on RF Module" , 2014

[4] http://igin.com/article-218-drip-irrigationa-water-conserving-solution.html

Autonomic Cloud Resource Management

Bhupesh Kumar Dewangan
University of Petroleum and Energy Studies
Dehradun, India
bhupesh.dewangan@gmail.com

Amit Agarwal
University of Petroleum and Energy Studies
Dehradun, India
aagarwal@ddn.upes.ac.in

Venkatadri M
University of Petroleum and Energy Studies
Dehradun, India
venkatadri.mr@gmail.com

Ashutosh Pasricha
Schlumberger: Oilfield Services
New Delhi, India
apasricha@slb.com

***Abstract*— Resource utilization of cloud affect the operational cost of cloud services. Since cloud user and demands increasing exponentially, the service provider needs to manage the recourse accordingly so that maximum profit can be provide to the service provider as well as cloud user with the quality of service constraint (QoS). To maintain QoS, service level agreement (SLA) violation rate, energy consumption by resources, cost, and execution time should be less. The energy efficiency and SLA violation rate are the major focused key point of this work. In this paper, energy consumption has been reducing through self-optimization, and SLA violation rate is minimized by self-healing methods and separate faulty VM from the resource pool. In continue, the operating cost of resources has been optimizing and less execution time has recorded. The proposed method is simulated in cloudsim toolkit, evaluates the performance metrics with a different set of workloads and the observation of this research and its experimental results and comparative analysis with existing frameworks are evidence of utmost performance.**

Keywords—Energy-optimization, Fault-tolerance, SLA violation rate, Resource Cost, Performance.

I. INTRODUCTION

Adaptability is one of the oversee highlights required in Cloud planning which is accomplished through changed alteration in resource provisioning in light of the solid changes in the client's workload and arrangements. Task booking has created as one the obsession in dispersed figuring since inefficient errand arranging can incite execution degradation and break of Service-Level-Agreement (SLA). In this way, feasible booking figuring's are required to restrict both estimation based estimations, for example, response time, structure throughput, system utilize and framework based estimations, for instance, movement volume, orchestrate correspondence cost and data correspondence cost.

The organization approach in the cloud structure is the ideal approach to manage to pick the operational cost, client verification strategy, and the nature of the association. Considering a given plan of clients having stamped appropriate SLAs with the cloud ace alliance, the preferred standpoint affiliation issue in the cloud structure can be, portrayed as the issue of streamlining any of the satisfactorily showed target limits subject to the given SLAs. The preferred standpoint affiliation decisions consolidate doling out VMs to servers, circling resource for each VM and moving them between servers to address SLA encroachment, top control necessities or warm emergencies.

To adjust to the Cloud stream, watching and reconfiguration frameworks might be an offer by cloud providers. Checking contains in lighting up captivated parts about the status of an advantage or an organization. While in the meantime, reconfiguration is a runtime change of the structure or the utilization of a foundation. As incalculable heterogeneous assets are secure with setting up a network and cloud system, this expands the disappointment and security concerns. The heterogeneity and dynamicity of the appropriated figuring condition additionally frustrate the issue. The utilization of scattered enlisting for work process applications is in this manner troublesome and there exist not a lot of endeavors to use IaaS hazes for such applications.

Autonomic resource management is an ability to improve use of benefits and customer satisfaction in autonomic structures, which are self-healing, self-configuring and self-optimizing. Self-healing is a limit of an adroit structure to recognize, look at and recover from grievous deficiencies thusly. Self-configuring is a capacity of an insightful structure to change in accordance with the modifications in nature. Self-Optimizing is the capacity to capably help resource appropriation and use for satisfying essentials of different customers. Given these characteristics, programming structures can be modernizing to take healing exercises if endeavors are curbing by a fault or frustration is distinguished. The therapeutic exercises can fuse changes to setups, recovering a technique or application that has failed, guarantee a section over-load that would cause a bottleneck in the work procedure, and streamline system execution.

II. RELATED WORK

In this section, the existing frameworks on self-optimization and self-healing have been explained in brief.

A. Self-optimization based Resource Management

Authors [1], present energy optimization, which intends to help resource use and unequivocally considers both dynamic and sits out of apparatus imperativeness use, which maximizes the resource utilization and minimizes the operating cost. [2] Is general and goes past the current state

of the quality by constraining both the number of movements required for mix and asset use in a single count with a course of action of considerable differences and conditions. The paper [3], with respect to disseminated registering, for the most part, centers on execution, cost and execution time of asset planning approaches.

Authors [4] shown that the execution of the proposed TROA is the best figuring in perspective of the parameters of slightest or low power use of assets with the high utilization, minimum cost, better makespan, and less CO2 outpouring prompts condition reasonability. Where, [5] model executed in cloudsim toolbox. One of the heuristics prompts basic diminishing of the vitality use by a Cloud server to develop by 83% curiously with a non-control cautious structure and by 66% on the other hand with a framework that applies just DVFS system yet does not modify fragment of VMs in run-time.

In [6], highlights the piece of correspondence surface and exhibits an arranging game plan, named e-STAB, which considers development necessities of cloud applications giving imperativeness successful occupation assignment and action stack modifying in server cultivate frameworks. In the paper, [7] their arranging count can profitably construct asset use; thus, it can decrease the asset usage for executing businesses. The test occurs to exhibit that their arrangement can diminish more asset usage than various plans do.

[8] Look at the association between establishment sections and power use of the disseminated figuring condition and discuss the planning of task forms. The paper [9] perform diversions with two load takes after. The outcomes demonstrate that the PP20 mode can set aside to 46.3% of power use with a drop rate of 0.03% on one load take after, and a drop rate of 0.12% with a power diminishment rate of 46.7% on the other. Cloud RAN [10] is another consolidated perspective in perspective of virtualization advancement that has created as a promising plan and capably keeps an eye on such issues. C-RAN outfit's high essentialness efficiency together with gigabit-per-second data rates transversely finished programming described remote frameworks.

[11] EnaCloud is a vitality sparing application live position approach for the vast size of the cloud stage. A vitality mindful heuristic calculation is proposed to pick a suitable blueprint for dynamic application arrangement. [12] Particularly base on the island parallel model and the multi-start parallel illustrate. Their new technique relies on intense voltage scaling (DVS) to restrict imperativeness use. [13] Give a centrality able extraordinary offloading and resource-provisioning plan to lessen significance utilize and condense application-finishing time.

[14] Play out extensive observations of a structure supporting the appropriated processing perspective as for imperatives capability. Soccer [15] create an energy-aware cloud resource management, which minimizes the energy consumption by using the self-optimization. Resources are schedule by the threshold value and fulfill the QoS. Chopper [16] the autonomic qualities are executed in the genuine cloud by utilizing fuzzy and machine learning ideas. The workloads are classified into various groups as per it's composed.

B. Self-healing based Resource Management

AFTRC [17] instrument is an adaptation to non-critical failure e-demonstrate for continuous a cloud and virtualized show which endures the blame proactively on the premise handling hub and unwavering quality. SRFSC [18] proposed a proactive acclimation to interior frustration plot for cloud applications using programming recuperation framework. Each cloud application is considered as a mix of interconnected cloud advantage parts, which may be perseveringly either couple or be around the couple. These parts may pass on finished a brisk LAN by systems for the remote method call.

Presents autonomic cloud resource administration by vitality and fault-tolerance techniques. [19]FTCloud is a system, which decides faults naturally. FTDG [20] proposed a fault-tolerance organizing structure using preemptive migration for stream figuring. The structure building joins four working spaces, customer space, chart space, storm space, and gear space. CBFIT_PFT [21] is best in class for inner faults of VM's. This proactive framework for non-critical fault-tolerance is beating by the amendments when the accuracy plunges under 20%.

In addition, FTWS [22], proposes a structure to compute cloud resource usage in two different ways, on spot, and on sales to diminish the cost of execution. This accuse tolerant method finds the cumbersome end of spot cases and execution cost as well. Tests demonstrate that it accomplishes 70% outcomes to diminish the execution cost.

III. PROBLEM STATEMENTS

The state of art survey of energy optimization and fault-tolerance are mentioning some findings in this section that can be improve for better performance.

A. SLA violation rate can be minimized,
B. Scope to minimizes the execution time,
C. The resource utilization can be maximized,
D. fault-tolerant techniques can be more accurate to identify a faulty resource,
E. The resource cost is more and the operational cost is high,
F. Energy consumption can be optimize.

IV. METHODOLOGY

The cloud user submits the workloads to the task manager for asking a different kind of services from the cloud server. The workloads are sorting in the order according to its priority, here execution time and the cost is taking as a priority of workloads. The following workloads are considering VM (Resource) allocation, which is showing in table 1.

TABLE I. WORKLOAD DATASET [15]

Workload	QoS requirements
Websites	Reliable storage, high network bandwidth, high availability
Technological computing	Computing capacity, reliable storage
Endeavour software	Security, high availability, customer confidence level, correctness
Performance testing	Computing capacity, network bandwidth, latency
Online transaction processing	Security, high availability, internet accessibility, usability
E-com	Variable computing load, customizability
Central financial services	Security, high availability, changeability, integrity
Storage and backup services	Reliability, persistence
Productivity applications	Network bandwidth, latency, data backup, security
Software/project development and testing	User self-service rate, flexibility, creative group of infrastructure services, testing time
Graphics oriented	Network bandwidth, latency, data backup, visibility
Critical internet applications	High availability, serviceability, usability
Mobile computing services	High availability, reliability, portability

A. The architecture of Proposed Method

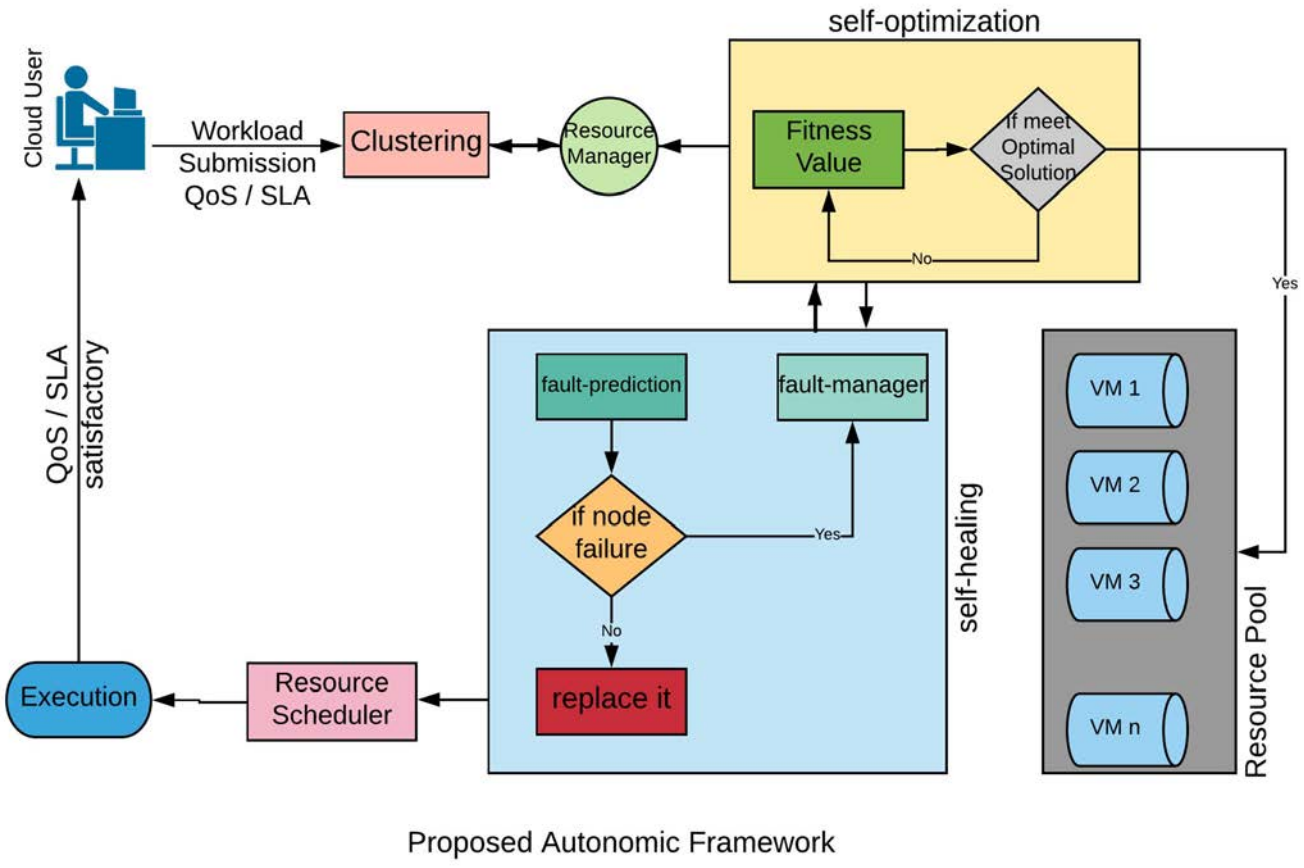


Fig.1. The Architecture of Proposed Autonomic Resource Scheduler Framework

B. Algorithm for energy-consumption

*Algorithm*1: Energy Consumption

1. Start
2. Initialize totalEnergy=0.0, max=1, k=0.5, and Antlion Algo.
3. getPower(utilization)
4. Compute s=k*max+(1-k)*max*utilization
5. End

C. Agorithm for Resource Utilization

*Algorithm*2: Resource Utilization

1. Start
2. Initialize maxiteration = 100, and Antlion Algo.
3. if(i==Population)
4. compute cpu_u =(Vm_Mips.get(i)/Pm_Mips);
5. bw_u =(Vm_BandWidth.get(i)/Pm_Bw);
6. ram_u =(Vm_Ram.get(i)/Pm_Ram);
7. tot=(cpu_u+bw_u+ram_u);
8. End if
9. End

D. Algorithm for Resource Scheduling

Algorithm 3: Assigning the workloads to VM's

1. Start
2. If MaxIter reached
3. For W= n to 1 // n is the number of VMs
4. Choose a nth task from the sorted list
5. Compute fitness function
6. Assign nth task to selected VM
7. Else, Update the ant lions position randomly. And Perform the ant lion algorithm
8. Return as the solution, the solution attributable to the most expert ant
9. End
10. The primary subterranean Ant in the lion subterranean Ant is mapped to every one of the mappings performed in stage 27
11. Mapping alternate ants in the insect lion calculation, with arbitrary task of every single existing assignment to various kinds of apparatus accessible
12. End

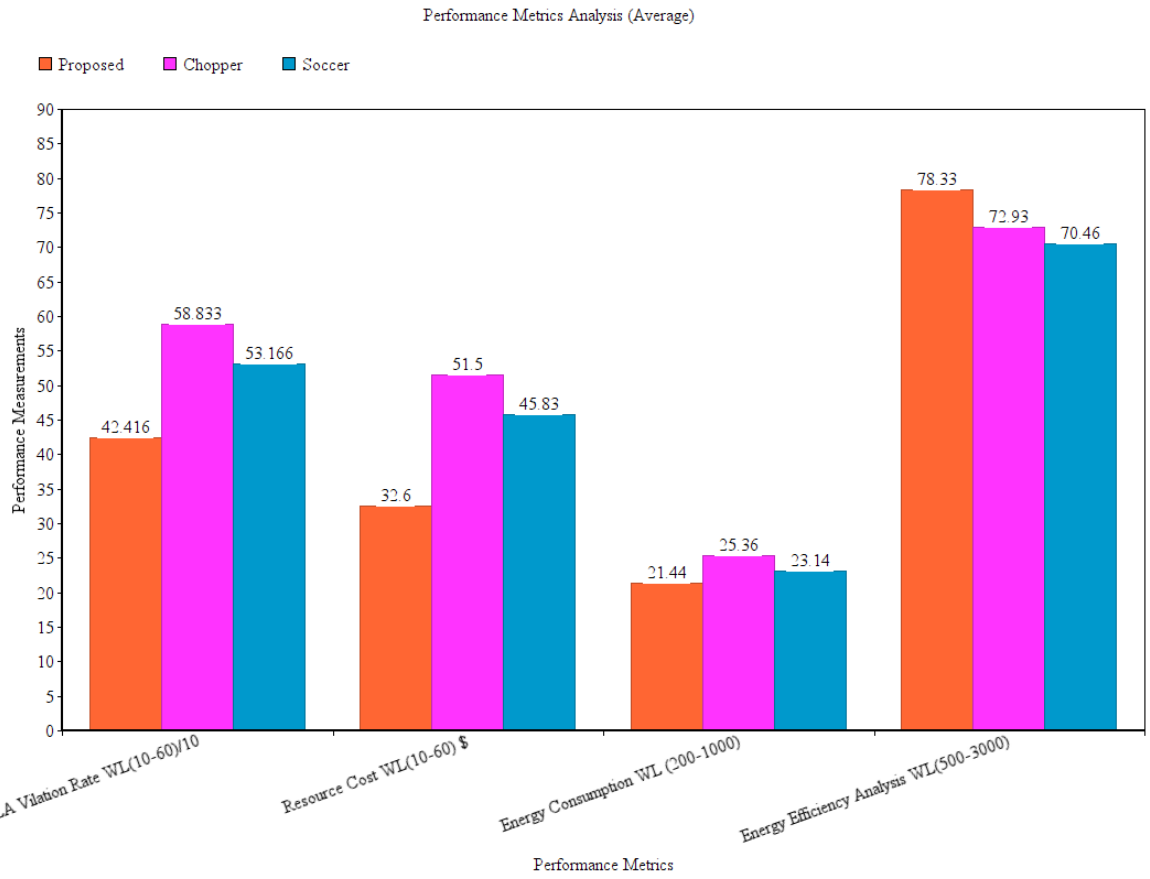


Fig. 2. Analysis of SLA Violation Rate, Resource Cost, Energy the Consumption and Energy-efficiency

V. Results and Analysis

The objective is to find the best VM to allocate the workloads submitted by a cloud user. Initially, 20 VM and 100 workloads submitted and simulated. The objective is to maximize the VM utilization, so best VM need to be identified, here the best VM identified by the fault-tolerant technique by identifying and rejecting faulty VM from the resource pool. The best VM value is identified by computing its energy consumption, SLA violation rate, and its utilization. Utilization value is computing by adding RAM, CPU and Bandwidth utilization of each VM. Below threshold, VM value has been assumed as faulty VM. The energy consumption is computing by algorithm1. Now both workloads and VMs are ready to assign. By utilizing algorithm3 VM's are assigned to workloads submitted by cloud user and resource utilization is compted by algorithm2. The proposed work is simulating in cloudsim toolkit and results are comparing with two autonomic existing model soccer and chopper. The detailed analysis is discussed in the following section. The energy consumption rate is increasing while increasing the workloads to resources. The experiments are testing on 200, 400, 600, 800 and 1000 workloads. The average results of each method are plotting in fig.1 and the results of the proposed work if performing better. When the energy is optimizing the following performance measurement has been taken to evaluate the performance of the proposed model and compared with the other two existing model.

A. Maximizes Energy-Efficiency Rate

The energy efficiency rate of the proposed model is computing while increasing the workloads from 500 to 3000. The efficiency rate of the proposed model and the other two models are computing and taking the average value for plotting the graph. The average value of the energy efficiency rate of proposed work is better.

B. Minimizes SLA Violation Rate

The SLA violation rate is computing for workloads 10 to 60. The average value of the violation rate is plotting for proposed and the other two existing model, and observe that the proposed model is having less SLA violation rate.

C. Minimizes Resource Cost

The resource cost of the proposed model and others model is computing for workloads 10 to 60, and it is observed that the proposed model producing resources to workloads at low cost. The average resource cost value of the proposed model and other models are observing and plotting in the graph.

The above analysis of energy-efficiency, SLA violation rate, and resource cost are plotting in graph fig.2. by its average values. In all, the above three analysis presents that the proposed model is having better performance. All the performance metrics computing and testing in different workloads, which is present in the table.

VI. Conclusion

Operating cost of any service provider is a major concern of cloud service provider and cloud user as well. In this work, two key parameters are considering for minimizing the cost of the resource, one is energy consumption and second is fault-tolerance for identifying faulty VM's. The VM's are separated from the resource pool based on best VM value, which is directly proportional to cost and in addition, the CPU, RAM and Bandwidth utilization of each VM's are computing to evaluate VM utilization. The best VM are separate for assigning the workloads submitted by a cloud user. The analysis of the proposed work is presenting the best performance in respect to cost, SLA violation rate, and energy-efficiency. The workloads are not filtering in this work to find the malicious workloads, if the workloads are submitting to the task manager then VM will also assign to malicious workloads that may cause of degradation of system performance. In future work, the self-protection method will consider to identifying and separating the malicious workloads from task manager.

References

[1] Lee, Y. C., & Zomaya, A. Y. (2012). Energy efficient utilization of resources in cloud computing systems. *The Journal of Supercomputing*, *60*(2), 268-280.

[2] Ghribi, C., Hadji, M., & Zeghlache, D. (2013, May). Energy efficient VM scheduling for cloud data centers: Exact allocation and migration algorithms. In *Cluster, Cloud and Grid Computing (CCGrid), 2013 13th IEEE/ACM International Symposium on* (pp. 671-678). IEEE.

[3] Berl, A., Gelenbe, E., Di Girolamo, M., Giuliani, G., De Meer, H., Dang, M. Q., & Pentikousis, K. (2010). Energy-efficient cloud computing. *The computer journal*, *53*(7), 1045-1051.

[4] Y. a. G. Alex. (2017). Comparison of Resource Optimization Algorithms in Cloud Computing, *International Journal of Pure and Applied Mathematics*, pp. 847-854,

[5] Beloglazov, A., & Buyya, R. (2010, May). Energy efficient resource management in virtualized cloud data centers. In *Proceedings of the 2010 10th IEEE/ACM international conference on the cluster, cloud and grid computing* (pp. 826-831). IEEE Computer Society.

[6] Kliazovich, D., Arzo, S. T., Granelli, F., Bouvry, P., & Khan, S. U. (2013, August). e-STAB: Energy-efficient scheduling for cloud computing applications with traffic load balancing. In *Green Computing and Communications (GreenCom), 2013 IEEE and Internet of Things (iThings/CPSCom), IEEE International Conference on and IEEE Cyber, Physical and Social Computing* (pp. 7-13). IEEE.

[7] Wu, C. M., Chang, R. S., & Chan, H. Y. (2014). A green energy-efficient scheduling algorithm using the DVFS technique for cloud datacenters. *Future Generation Computer Systems*, *37*, 141-147.

[8] Luo, L., Wu, W., Di, D., Zhang, F., Yan, Y., & Mao, Y. (2012, June). A resource scheduling algorithm of cloud computing based on energy efficient optimization methods. In *Green Computing Conference (IGCC), 2012 International* (pp. 1-6). IEEE.

[9] Duy, T. V. T., Sato, Y., & Inoguchi, Y. (2010, April). Performance evaluation of a green scheduling algorithm for energy savings in cloud computing. In *Parallel & Distributed Processing, Workshops and Phd Forum (IPDPSW), 2010 IEEE International Symposium on* (pp. 1-8). IEEE.

[10] Pompili, D., Hajisami, A., & Tran, T. X. (2016). Elastic resource utilization framework for high capacity and energy efficiency in cloud RAN. *IEEE Communications Magazine*, *54*(1), 26-32.

[11] Li, B., Li, J., Huai, J., Wo, T., Li, Q., & Zhong, L. (2009, September). Enacloud: An energy-saving application live placement approach for cloud computing environments. In *2009 IEEE International Conference on Cloud Computing* (pp. 17-24). IEEE.

[12] Mezmaz, M., Melab, N., Kessaci, Y., Lee, Y. C., Talbi, E. G., Zomaya, A. Y., & Tuyttens, D. (2011). A parallel bi-objective hybrid metaheuristic for energy-aware scheduling for cloud computing systems. *Journal of Parallel and Distributed Computing*, *71*(11), 1497-1508.

[13] Guo, S., Xiao, B., Yang, Y., & Yang, Y. (2016, April). Energy-efficient dynamic offloading and resource scheduling in mobile cloud computing. In *INFOCOM 2016-The 35th Annual IEEE International Conference on Computer Communications, IEEE* (pp. 1-9). IEEE.

[14] Mastelic, T., Oleksiak, A., Claussen, H., Brandic, I., Pierson, J. M., & Vasilakos, A. V. (2015). Cloud computing: Survey on energy efficiency. *Acm computing surveys (csur)*, *47*(2), 33.

[15] Singh, S., Chana, I., Singh, M., & Buyya, R. (2016). SOCCER: self-optimization of energy-efficient cloud resources. *Cluster Computing*, *19*(4), 1787-1800.

[16] Gill, S. S., Chana, I., Singh, M., & Buyya, R. (2017). CHOPPER: an intelligent QoS-aware autonomic resource management approach for cloud computing. *Cluster Computing*, 1-39.

[17] Malik, S., & Huet, F. (2011, July). Adaptive fault tolerance in real time cloud computing. In *Services (SERVICES), 2011 IEEE World Congress on* (pp. 280-287). IEEE.

[18] Liu, J., Zhou, J., & Buyya, R. (2015, June). Software rejuvenation based fault tolerance scheme for cloud applications. In *Cloud Computing (CLOUD), 2015 IEEE 8th International Conference on* (pp. 1115-1118). IEEE.

[19] Zheng, Z., Zhou, T. C., Lyu, M. R., & King, I. (2010, November). FTCloud: A component ranking framework for fault-tolerant cloud applications. In *2010 IEEE 21st International Symposium on Software Reliability Engineering* (pp. 398-407). IEEE.

[20] Sun, D., Zhang, G., Wu, C., Li, K., & Zheng, W. (2017). Building a fault tolerant framework with deadline guarantee in big data stream computing environments. *Journal of Computer and System Sciences*, *89*, 4-23.

[21] Sampaio, A. M., & Barbosa, J. G. (2017). A comparative cost analysis of fault-tolerance mechanisms for availability on the cloud. *Sustainable Computing: Informatics and Systems*.

[22] Poola, D., Ramamohanarao, K., & Buyya, R. (2014). Fault-tolerant workflow scheduling using spot instances on clouds. *Procedia Computer Science*, *29*, 523-533.

Comparative Analysis between Proprietary Software VS. Open-Source Software VS. Free Software

Abhineet Anand
SCSE, Galgotias University
Graeter Noida, India
abhineet.mnnit@gmail.com

Anand Krishna
Computer Science Dept
SRM University, India

Rajeev Tiwari
School of Computer Science,
UPES,Dehradun, India

Robin Sharma
Students, SCSE
Galgotias University
Graeter Noida
robinsharma3341@gmail.com

Abstract*—This analysis mainly focuses on the domain of prime importance of Software which are more prominent for our system. It requires more attention of ours to switch over in Open-Source rather to invest and compete with proprietary software. In this model, we have considered a seesaw model which balances open-source software and proprietary software by a balance beam of free software. We have taken in consideration two fields (i.e) competition and investment to compare the consumer's interest areas and to know about their choices in current scenario. On putting loads of consumers on one side we can also recorded that the beam balance (free software) moves automatically towards the adjacent software and enrich the given software, in both fields whether it is related to competition or in terms of consumer's investment. This model analysis on points of fully covered market and partly covered market, which shows the current trends of consumer's interest and their profit maximization policies.

Keywords*— Open-Source Software, Proprietary Software, Free Software, Competition, Investment.

I. INTRODUCTION

The term Open-Source originated in the context of software development to designate a specific approach to creating computer programs. Today however Open-source designates a broader set of values what we call the open-source way. Open-source projects, products or initiatives embrace and celebrate principles of open exchange, collaborative participation, rapid prototyping, transparency, meritocracy and community-oriented development[1].

There are several ways in which work on an open-source project can start:

(1.) An individual who senses the need for a project announces the intent to develop a project in public.
(2.) A developer working on a limited but working code-base, releases it to the public as the first version of an open-source program.
(3.) The source code of a mature project is released to the public.
(4.) A well-established open-source project can be forked by an interested outside party.
Eric Raymond observed in his essay "The Cathedral and Bizarre" that announcing the intent for a project is usually inferior to releasing a working project to the public[2][13].
It's a common mistake to start a project when contributing to an existing similar project would be more effective (NIH Syndrome). To start a successful project it is very important to investigate what's already there. The process starts with a choice between the adopting of an existing project, the starting of a new project [2]. If a new project is started, the process goes to the Initiation phase. If an existing project is adopted, the process goes directly to the Execution phase.

Free Software is software that can be freely used, modified and redistributed with only one restriction: any redistributed version of the software must be distributed with the original terms of free use, modification and distribution (known as copy left).The definition of free software is stipulated as part of the GNU Project and by the Free Software Foundation. Free Software may be packaged and distributed for a fee; the "free" refers to the ability to reuse it, modified or unmodified, as part of another software package[3][6]. AS part of the ability to modify, users of free software may also have access to and study the source code. The concept of free software is the brainchild of Richard Stallman, head of the GNU Project. The best known example of free software is Linux, an operating system that

is proposed as an alternative to Windows or other proprietary operating systems.

The first formal definition of free software was published by FSF in February 1986.That definition, written by Richard Stallman, is still maintained today and states that software is free software if people who receive a copy of the software have the following four freedoms.The numbering begins with zero, not only as a spoof on the common usage of zero-based numbering in programming languages, but also because "Freedom 0" was not initially included in the list, but later added first in the list as it was considered very important.

- Freedom 0: The freedom to run the program for any purpose.
- Freedom 1: The freedom to study how the program works, and change it to make it do what you wish.
- Freedom 2: The freedom to redistribute and make copies so you can help your neighbor.
- Freedom 3: The freedom to improve the program, and release your improvements (and modified versions in general) to the public, so that the whole community benefits.

Freedoms 1 and 3 require source code to be available because studying and modifying software without its source code can range from highly impractical to nearly impossible[4][5].

Proprietary software is software that is owned by an individual or a company (usually the one that developed it). There are almost always major restrictions on its use, and its source code is almost always kept secret. Source code is the form in which a program is originally written by a human using a programming language and prior to being converted to machine code which is directly readable by a computer's CPU (central processing unit). It is necessary to have the source code in order to be able to modify or improve a program. Software that is not proprietary includes free software and public domain software. Free software, which is generally the same as open source software, is available at no cost to everyone, and it can be used by anyone for any purpose and with only very minimal restrictions. These restrictions vary somewhat according to the license, but a typical requirement is that they include a copy of the original license. The most commonly used license, the GNU Public License (GPL), additionally requires that if a modified version of the software is distributed, the source code for such modified version must be made freely available. The best known example of software licensed under the GPL is Linux. Public domain software is software that has been donated to the public domain by its copyright holder. Thus it is no longer copyrighted. Consequently, such software is completely free and can be used by anybody for any purpose without restriction.

II. LITERATURE SURVEY

It is been investigated from decade towards our need and switch over towards open source. This widely asked question from decade on open source software is to identify its competition and investment strategies in market for developers and consumers to give their contribution to open source software [4][7][8].

Our see-saw model focuses upon different compatibility strategies of open source software and contrast the competition and investment between open source software and proprietary software. It is been concluded that open source software is not necessarily inferior in quality to proprietary software[6][7].

This research took consideration on two different markets that were fully covered market and partly covered market. The two fields that were taken in notice were competition between open source and proprietary software. In the litrature by different renowned scientist the concept has been modeled which disribed the competition between Windows and Linux as a dynamic “mixed duopoly”, where a not-for-profit competitor interacts with a for-profit competitor[10]. The second field talks about the part of investment of consumers in the market where open-source software are seen in good position as due to their less cost of money as compared with the Proprietary software.

The different research paper takes approach of consumers in the market and plotted them in a see-saw model. This helps to characterize and correlate the needs of ours to basically move towards which software, this is been explained through this model. We seek to find the best compatibility strategy software through this model and the taste difference of consumers in the market through competition and investment[11].

This model has homogeneous division of consumers on both ends. The consumers’ preferences for products/services are been recorded, which depends upon factor, the cost of adopting a software product for consumers and their past experiences with the product[12-13].

Due to this factor, the consumers have different tastes for the products. The consumers who have lower taste for the open source software would rather choose the proprietary software even though the open source software is free of charge.

The difference between consumers taste allows interpretations for real-world software competition and investment in the market.

III. BASIC MODEL

Consider a software market in a form of see-saw in which products are located at the ends of seesaw and free software as the mid-point of see-saw balancing the two products (i.e) on extreme left open-source software(O) at s=0 and on extreme right Proprietary Software (P) at s=l. The consumers are uniformly distributed along both the sides of it so they balance each other with free software.

Consumers also might differ due to their different taste for products. Here, for a consumer located at € [0, l], incurs utility cost st if he uses the open-source software because of the difference between the ideal preferences and product specifications[6]. Similarly, it incurs utility cost t (1-s) if goes with the proprietary software, where t measures the consumers taste difference. Assume that the marginal costs of both the open-source and proprietary software products are zero. Since, free software lies at centre of see-saw and equals both the ends position and distance. So, if we consider length of see-saw as s then (O) distance from it becomes s whereas (P) distance from it becomes (1-s).

In this model, we assume that the two products have same inherent quality s and are incompatible, and covers the whole market (i.e.) all the consumers choose to use one of the two products to equal the Free software or unequal it(vice-versa). This would be true when the benefit of the product is sufficiently large.

If a consumer located at (O) adopts the open-source software and her utility U_o = a+ k*q –t*x+ d*q, where q is number of open source and k is the degree of contribution of each consumer to the quality of the open source software. The parameter d refers to the network externality that a software user receives from other users of same or compatible software. We assume that the open source software product is freely available, and there is no price component in its net utility. Similarly, if the consumer located at (P) adopts the proprietary software, her net utility Up=a-t (1-s) +d*q*p -p, where q*p denotes the number of proprietary software users and p is the price of the proprietary software[7].

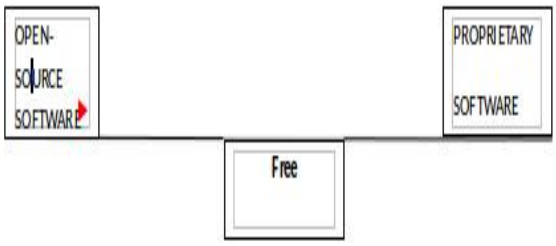


Figure 1: The Basic See-Saw Model

Suppose the consumer at S€[0,l] is indifferent between the open source and proprietary software products, then from Uo=Up, we have;
At equilibrium, (i.e) when free software balances the two of them.

$$x_c = \frac{p+t-\gamma}{2t-2\gamma-k}, q_o = \frac{p+t-\gamma}{2t-2\gamma-k}, q_p = 1-\frac{p+t-\gamma}{2t-2\gamma-k}.$$

on for profit maximization,

The profit for the proprietary software producer is:

$$\pi(p) = pq_p = p(1-x_c) = p[1-\frac{p+t-\gamma}{2t-2\gamma-k}]$$

By solving the profit maximization problem with respect to p, we get:

$$p^* = \frac{t-\gamma-k}{2}, \quad \pi^* = \frac{(t-\gamma-k)^2}{4[2(t-\gamma)-k]}$$

Where, * p and * π denote the equilibrium price and profit respectively.
Using *st to represent the market share of the proprietary software under the equilibrium price and *st to denote that of the open source software, we have

$$M_p^* = \frac{t-\gamma-k}{2[2(t-\gamma)-k]}, \quad M_o^* = \frac{3t-3\gamma-k}{2[2(t-\gamma)-k]}.$$

The base-level qualities of proprietary and open source software are identical. However, the quality of the open source software increases with the number of users, and the price of open source software is zero. Hence, in equilibrium, the open source software has a bigger market share than the proprietary software.

IV. COMPATIBILITY AND PROFITS ANALYSIS

In this basic model, we assumed that both open source and proprietary software were incompatible. Therefore, to make their product compatible users of proprietary software can move towards open-source software. This strategy of consumers helps them to excel in both fields of competition as well as investment. It also makes sure that the consumers if not satisfied with the services can also discontinue to it and that to without loosing of any cost of money.
Here, on basis of See-Saw Model, this analysis broadly depends upon two fields of consumers market, whether the market is fully covered or partly covered.

A. Fully Covered Market:

On analyzing the present utility functions of open-source software and proprietary software, the consumer's change according to different compatibility strategy when the market is fully covered is been analyzed and is been recorded on basis of competition and investment. Secondly, the net utilities of the open-source software and proprietary software consumers at equilibrium conditions are been calculated. Then on the basis of these outcomes we get the result under different compatibility strategies[8][9].

B. Partly Covered Market:

On analyzing the present utility functions of open-source software and proprietary software, the consumer's ratio drastically change according to compatibility strategy when the market is partly covered. This situation is been analyzed and is been recorded on the basis of competition and investment. Then the free software balanced beam is also took on consideration and been recorded. After on the

basis of these outcomes we get the result under different compatibility strategies.

V. CONCLUSION

From this see-saw model we have concluded the two fields of competition and investment in three different software fields that are Proprietary Software, Open-source software and Free Software. Here we also assume two different markets that were fully covered market and partly covered market. We have equally divided the consumers on both the ends of the see-saw model and balance them with Free Software. On applying equilibrium conditions and profit maximization policies we came to know about consumers interest the market with some survey and theoretical concepts. This is been noted that in fully covered market the dominance of Proprietary Software is been noted but in partly covered market scenario the Open-source Software and Free Software rules the market and have their supremacy whether it been in terms of investment or competition because these software's provides flexibility and consumers can earn benefits from them. They also have an option to discontinue with the program if they don't like, and that to without losing any cost of money. Therefore, this model helps us to understand the software's analysis and their balances with the free software.

REFERENCES

[1]. Soderberg, "J.,Hacking Capitalism", New York: Routledge, 2008, https://doi.org/10.4324/ 9780203937853

[2]. A. Singh, A. Anand. "Data leakage detection using cloud computing", International Journal Of Engineering And Computer Science, 6 (4), April 2017, .doi: https://dx.doi.org/10.18535/ijecs/v6i4.59

[3]. Anand, A., Sihag, V. K., & Gupta, S. N., "Wavelength conversion and deflection routing in all-optical packet-switched networks through contention resolution: A survey." In Proceedings of the cube international information technology conference, pp. 155–159, 2012, http://doi.acm.org/ 10.1145/2381716.2381747

[4]. Bessen, J., "What good is free software. Government policy toward open source software", pp. 12- 27, 2002.

[5]. Dumka, A., Mandoria, H. L., Dumka, K., & Anand, A, MPLS VPN using ipv4 and ipv6 protocol. In proceeding of 2nd international conference on computing for sustainable global development, p.p. 1051-1055, 2015.

[6]. Kuan, K. K., & Chau, P. Y., "A perception-based model for edi adoption in small businesses using a technology organization environment framework. Information & management", 38(8), 507–521.

[7]. Lakhani, Karim R. and Wolf, Robert G., "Why Hackers Do What They Do: Understanding Motivation and Effort in Free/Open Source Software Projects", 4425-03, September 2003.

[8]. Lerner, J., & Tirole, J. Some simple economics of open source. The Journal of Industrial Economics, 50(2), 197–234. 2002.

[9]. Meng, Z., Lee, S.-Y. T., "Open source vs. proprietary software: Competition and compatibility", 2005.

[10]. Singh, S., Anand, A., & Tiwari, R, "Optimization of data centres for heat management". In Intelligent communication, control and devices, Singapore: Springer Singapore, pp 1025-1032, 2018.

[11]. Cremer H., Marchand M., Thisse J.F, "Mixed oligopoly with differentiated products. Internat", J. Indust. Organ, 1991.

[12]. Dolley A., "IBM claims SCO conspiring with Microsoft over Linux. ZDNet Australia", September 2003. http://www.zdnet.co.nz/newstech/enterprise/story

[13]. Eric Raymond, "The Cathedral and the Bazaar" by David Wiley is licensed under a Creative Commons Attribution 4.0 International License.

Customer Segmentation using K-means Clustering

[1]Tushar Kansal, [2]Suraj Bahuguna , [3]Vishal Singh, [4]Tanupriya Choudhury

[1]kansaltushar18@gmail.com;[2]bahugunasooraj@gmail.com;[3]vishalsinghpari.0816@gmail.com,[4]tanupriya1986@gmail.com

[1,2,3,4] University of Petroleum & Energy Studies (UPES),Dept. of Informatics,School of Computer Science,Dehradun

Abstract-:

The zeitgeist of modern era is innovation, where everyone is embroiled into competition to be better than others. Today's business run on the basis of such innovation having ability to enthral the customers with the products, but with such a large raft of products leave the customers confounded, what to buy and what to not and also the companies are nonplussed about what section of customers to target to sell their products. This is where machine learning comes into play, various algorithms are applied for unravelling the hidden patterns in the data for better decision making for the future. This elude concept of which segment to target is made unequivocal by applying segmentation. The process of segmenting the customers with similar behaviours into the same segment and with different patterns into different segments is called customer segmentation. In this paper, 3 different clustering algorithms (k-Means, Agglomerative, and Meanshift) are been implemented to segment the customers and finally compare the results of clusters obtained from the algorithms. A python program has been developed and the program is been trained by applying standard scaler onto a dataset having two features of 200 training sample taken from local retail shop. Both the features are the mean of the amount of shopping by customers and average of the customer's visit into the shop annually. By applying clustering, 5 segments of cluster have been formed labelled as Careless, Careful, Standard,Target and Sensible customers. However, two new clusters emerged on applying mean shift clustering labelled as High buyers and frequent visitors and High buyers and occasional visitors.

Keywords: Customer Segmentation, k-Means algorithm, Mean shift algorithm, Agglomerative algorithm, Machine learning, Python.

Introduction:

As more and more business being coming up every day, it has become significantly important for the old businesses to apply marketing strategies to stay in the market as the competition has been cut to throat. Change or die have become the simple rule of marketing in today's world. As the customer base is increasing day by day it has become challenging for the companies to cater to the needs of each and every customer, this is where Data mining serves a very important role to unravel hidden patterns stored in the company's database. Customer segmentation is one of the application of data mining which helps to segment the customers with similar patterns into similar clusters hence, making easier for the business to handle the large customer base. This segmentation can directly or indirectly influence the marketing strategy as it opens many new paths to discover like for which segment the product will be good, customising the marketing plans according to the each segment, providing discounts for a specific segment, and decipher the customer and object relationship which has been previously unknown to the company. Customer segmentation allows companies to visualise what actually the customers are buying which will prompt the companies to better serve their customers resulting in customer satisfaction, it also allows the companies to find who their target customers are and improvise their marketing tactics to generate more revenues from them.

Clustering has been proven effective to implement customer segmentation. Clustering comes under unsupervised learning, having ability to find clusters over unlabelled dataset. There are a number of clustering algorithm over which like k-means, hierarchical clustering, DBSCAN clustering etc. In this paper, three different clustering algorithms have been implemented over a dataset with two features with 200 records.

K-means Clustering:

It is the simplest algorithm of clustering based on partitioning principle. The algorithm is sensitive to the initialization of the centroids position, the number of K (centroids) is calculated by elbow method (discussed in later section), after calculation of K centroids by the terms of Euclidean distance data points are assigned to the closest centroid forming the cluster, after the cluster formation the barycentre's are once again calculated by the means of the cluster and this process is repeated until there is no change in centroid position.[10][11]

Agglomerative Clustering-:

Agglomerative Clustering is based on forming a hierarchy represented by dendrograms (discussed in later section). Dendrogram acts as memory for the algorithm to tell about how the clusters are being

formed. The clustering starts with forming N clusters for N data points and then merging along the closest data points together in each step such that the current step contains one cluster less than the previous one.

Mean shift Clustering-:

This clustering algorithm is a non-parametric iterative algorithm functions by assuming the all the data points in the feature space as empirical probability density function. The algorithm clusters each data point by allowing data point converge to a region of local maxima which is achieved by fixing a window around each data point finding the mean and then shifting the window to the mean and repeat the steps until all the data point converges forming the clusters.

Elbow Method-:

Elbow method is used for finding optimal value of K for K-means clustering algorithm. This method works by finding the SSE of each data point with its nearest centroid with different values of K. As value of K increases the SSE will decrease and at a particular value of K where there is most decline in the SSE is the elbow, the point at which we should stop dividing data further.

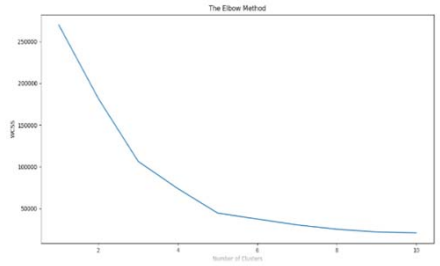


Fig. 1: Graph for Values of K VS WCSS(Within Cluster Sum of)

From the above graph it is clearly be seen that from number of cluster = 4 to number of cluster = 5 there has been substantial decrease hence, we choose the K value for our dataset as 5.

Dendrogram-:

Dendrogram is the hierarchical representation of object, it is used to determine the output of the hierarchical clustering. The way Dendrogram is interpreted is by checking the height of each clade (horizontal line), the lower the height the more associated data points are and greater the height more less associated data points.

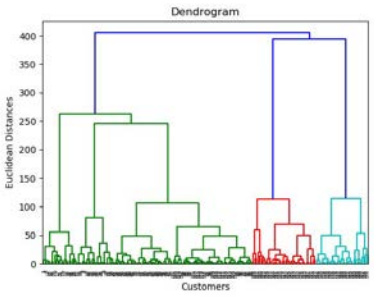


Fig2: Dendrogram Structure of the dataset.

Fig (2) shows Dendrogram of our dataset built using [9].The figure displays how the clusters are being formed to eventually converge to a single cluster. Dendrogram is being used used for finding the numberof clusters that is optimal to apply for Agglomerative clustering, in the Fig (2) we look for the longest vertical line which is not being cut by any of the clades (horizontal line) extended virtually over complete width of the graph, the second last clade of green colour have it's right leg bearing the longest vertical line which is not been cut by any clade. Now, by drawing a hypothetical horizontal line cutting through the longest vertical line, we get the horizontal line cutting total 5 vertical lines providing us the optimal number of clusters for our dataset.

Bandwidth:

Bandwidth can be considered as the radius of the circle (kernel) describing how much the data points should be in the cluster. It is the only input requisite for the Mean shift algorithm, calculated by the help of K Nearest Neighbours. Mean shift algorithm is very much sensitive to theinitialization of bandwidth, a small value can slow down the converging process while a large value can speed up convergence.

Methodology:

Data Collection:

The dataset has been taken from a local retail shop consisting of two features, average number of visits to the shop and average amount of shopping done on yearly basis.

Feature Scaling:

The data has been scaled using Standard Scaler [9], by applying standard scaler the data gets centred around 0 with standard deviation of 1.

$$\frac{x - mean(X)}{stdev(X)}$$

x = entry in a feature set xi ϵ X

mean (X) = mean of feature set X

stdev (X) = standard deviation of X

K means Clustering:

Choosing the optimal number of clusters:

Elbow method is applied to calculate value of K for the dataset.

Step-1: Run the algorithm for various values of k i.e making the k vary from 1 to 10.

Step-2: Calculate the within cluster squared error.

Step-3: Plot the calculated error, where a bent elbow like structure will form, will give the optimal value of clusters.

SSE is calculated by -:

$$\sum_{i=1}^{k} \sum_{Xj \,\epsilon\, Si} \|Xj - \mu i\|^2$$

X_j = data point in S_i cluster

μ_i = centroid of the cluster

Algorithm:

Step-1: Initialize the K (= 5) clusters.

Step-2: Assign the data point that is closest to any particular cluster.

Step-3: Recalculate the centroid position based on the mean of the cluster formed

Step-4: Repeat step 2 and 3 until the centroid position remains unchanged in the previous and current iteration.

Fig. 3: Clusters formed by K means

The figure above shows the 5 final clusters where the cluster in orange colour gives the target customer.

Agglomerative Clustering:

Choosing the optimal number of clusters:

Cluster value for this algorithm have been calculated by the Dendrogram as described in Dendrogram section which also gave the value of K = 5.

Algorithm:

1) Each data point is taken as to be a cluster.
2) Merge the two closest cluster.
3) Step 2 needs to be repeated until all the data points are merged together to form a single cluster. However, as we have defined the value of K as 5, the algorithm will stop when all the data points are part of any of the 5 clusters.

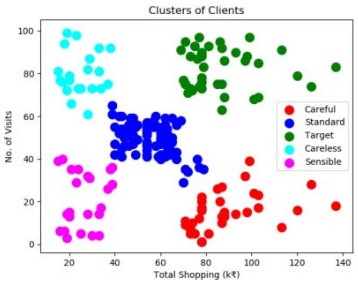


Fig.4: Clusters formed by Agglomerative Clustering

Since, the number of clusters for K-means and Agglomerative are equivalent they gave same pattern for final clusters. However, if you look closely a point on the top left corner in Standard cluster has changed its cluster and similar case is with 2-3 points on lower right corner in Standard cluster.

Mean Shift Clustering:

This non-parametric clustering method is being applied to see some different pattern in a dataset as K-means and Agglomerative gave almost the same result. There is no need of choosing the number of clusters. However, it needs one input parameter, bandwidth (radius) which is calculated using K-nearest neighbour algorithm. This algorithm follows an iterative approach where a point of local maxima is found around each data point defined by probability density function, and iterates until when all the data point converges up the hill (created by PDF), also known as 'hill climbing algorithm'.

PDF can be estimated by-:

$$\widehat{f(x)} = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

h = bandwidth

K = kernel

Some examples of Kernel:

1) *Rectangular:*

$$f(x) = \begin{cases} 1, & a \leq x \leq b \\ 0, & else \end{cases}$$

2) *Gaussian:*

$$f(x) = e^{\frac{-x^2}{2\sigma^2}}$$

Algorithm:

1) A window is associated around each data point created by PDF
2) Mean around the window is calculated
3) Window is moved towards the newly calculated mean.
4) Step 2 and 3 are repeated until when all the data points converge to a local maxima resulting in clusters.

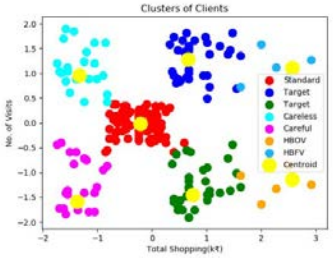


Fig. 5: Clusters formed by Mean Shift

The final outcome gave two new clusters labelled as High Buyer Frequent Visitors and High Buyer Occasional Visitors, these two new clusters can help the retail shop to treat the customers lying their segment as their VIP customers providing them accolade on purchase of products or giving them extra discounts.

Literature Review:

Customer Segmentation:

Jayant et al. [1] states that in customer segmentation, the customers are divided into different groups where customers of the same group are similar to each other in terms of marketing. Customers are divided into different clusters based on various attributes such as age ,interests, age, spending habits etc.

Sulekha et al. [7]provides the four popular bases for segmentation

1) Geographic Segmentation: segmentation on the basis geographic region, population density or climate.
2) Demographic Segmentation: market segment on the basis of age, size and family type, etc.
3) Psychographic Segmentation: segmentation based on customer's life style variables like interests, opinions, attitudes etc.
4) Behavioural Segmentation**:** segmentation is based on actual customer behaviourtowards products like brand loyalty, user status, readiness to buy etc.

Customer segmentation is based on based on the strategy called divide and conquer by utilising the advantage of segmentation the marketers can gain advantage over a particular segment and slowly can prevail over other marketers. Using market segmentation the marketers can focus more on customer relationship management which was not earlier possible with existing mass marketing tactics.

Clustering-:

Clustering is the process of grouping the information in the dataset based on some similarities. There are a number of algorithms which can be chosen to be applied on a dataset based on the situation provided. However, no universal clustering algorithm exists that's why it becomes important to opt for appropriate clustering techniques. Vaishali et al. [8]. In this paper, we have implemented three clustering algorithms using python scikit learn library [9].

K-Means Clustering:

K- means algorithm is one of the most popular partitioning clustering algorithm. This clustering algorithm depends on the centroid where each data point is placed on one of the K non overlapping clusters which are selected before running of the algorithm, Chinedu et al. [2]. The clusters formed corresponds to the hidden pattern in the data which gives the required information to help in decision making process.

Agglomerative Clustering:

This clustering comes under hierarchical clusters are formed based on some hierarchy. Hierarchical clustering is It is based on the concept that objects that are closer are more related to each other in comparison of the objects that are far from each other., T.Nelson et al. [3]. The main challenge of Hierarchical method is that once it undergoes split or merge operation it can never be undone. This challenge is profitable as it leads to smaller computation costs by not worrying about a combinatorial number of different choices. Yogita et al. [4]. There are two strategy in hierarchical clustering, first is top-down strategy also known as divisive clustering and second is bottom – up strategy also known as agglomerative clustering.

Agglomerative clustering process is generally slower than divisive clustering but allows more flexibility because it permits the user to supply any arbitrary similarity function defining what constitutes a similar cluster pair to merge together, Omar et al. [5].

Mean shift Clustering:

Sulekha et al. [7] defines this algorithm as a gradient ascent technique. In mean shift the local maxima of a density function is found from the given data samples that are discrete. It works with a search window that is positioned over a section of the distribution. The mean shift technique is used for real data analysis which is an application dependent tool and initially shape of data cluster is not assumed. This algorithm has wide applications in object detection, image segmentation.

Results:

We have taken two internal clustering measure, silhouette score and Calinski-Harabasz index.

Silhouette Score:

It is a way of measuring how well the data point has been clustered into the correct cluster.

First Step-:

a = Average distance between the centroid of a cluster and the data points embroiled into it.

Second Step:

b = Average distance between the data point and the closest cluster data points.

Third Step:

Silhouette Score = $\frac{b-a}{\max(b,a)}$

For the data point to be well grounded in its cluster 'b' needs to be large and 'a' needs to be small so that difference between the two is as large as possible.

'max (b,a)' is added to normalized the silhouette score. Higher the score better the data point belongs to that cluster.

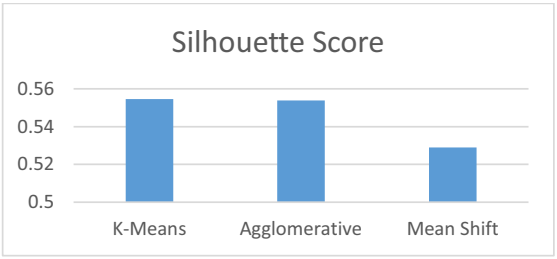


Fig. 6: Comparison of Silhouette Score

Figure above displays the silhouette score for the three algorithms applied in this paper, the graph shows there is not much significant difference in K-means and Agglomerative clustering. Hence, these two algorithms were able to cluster our data well than Mean shift algorithm as displayed by the low value of silhouette score.

Conclusion:

As our dataset was unlabelled, in this paper we have opted for internal clustering validation rather than external clustering validation, which depends on some external data like labels. Internal cluster validation can be used for choosing clustering algorithm which best suits the dataset and can correctly cluster data into its opposite cluster.

References:

[1] Jayant Tikmani, Sudhanshu Tiwari, Sujata Khedkar "Telecom customer segmentation based on cluster analysisAn Approach to Customer Classification using k-means", IJIRCCE,Year: 2015.

[2] Chinedu Pascal Ezenkwu, Simeon Ozuomba, Constance kalu Electrical/Electronics & Computer Engineering Department, University of Uyo, Uyo, Akwa Ibom State, Nigeria "Application of K-Means Algorithm for Efficient Customer Segmentation: A Strategy for Targeted Customer Services", IJARAI,Year: 2015.

[3] T.NelsonGnanarajDr.K.Ramesh Kumar N.Monica"Survey on mining clusters using new k-mean algorithm from structured and unstructured data", IJACST,Year: 2014.

[4] Yogita Rani and Dr. Harish Rohil"A Study of Hierarchical Clustering Algorithm", IJICT,Year: 2013.

[5] Omar Kettani, FaycalRamdani, BenaissaTadili"An Agglomerative Clustering Method for Large Data Sets", IJCA,Year: 2014.

[6] Snekha, ChetnaSachdeva, Rajesh Birok"Real Time Object Tracking Using Different Mean Shift Techniques–a Review", IJSCE,Year: 2013.

[7] SulekhaGoyat"The basis of market segmentation: a critical review of literature", EJBM,Year: 2011.

[8] Vaishali R. Patel and Rupa G. Mehta "Impact of Outlier Removal and Normalization Approach in Modified k-Means Clustering Algorithm", IJCSI,Year: 2011.

[9]Scikit-learn: https://scikit-learn.org

[10] Tanupriya Choudhury, Vivek Kumar, Darshika Nigam, Intelligent Classification & Clustering Of Lung & Oral Cancer through Decision Tree & Genetic Algorithm, International Journal of Advanced Research in Computer Science and Software Engineering,2015

[11] Tanupriya Choudhury, Vivek Kumar, Darshika Nigam, An Innovative and Automatic Lung and Oral Cancer Classification Using Soft Computing Techniques, International Journal of Computer Science & Mobile Computing,2015

Data Preservation by hash algorithm for matrix multiplication over venomous cloud

Gaurav Goel
Assistant Professor
CSE, CGC-COE
Landran, Mohali
gaurav.goel9@gmail.com

Rajeev Tiwari
Associate Professor
CSE, UPES
Dehradun.
rajeev.tiwari@ddn.upes.ac.in

Vinay Rishiwal
Associate Professor
MJP Rohilkhand Univ.
Bareilly, UP.
vianyrishiwal@gmail.com

Shuchi Upadhyay
Assistant Professor,
UCALS, Uttaranchal
Univ. Dehradun
Shuchi.diet@gmail.com

***Abstract*-In today Scenario various organizations, business are dependent on Cloud Computing. This paper focus on security issues in cloud computing. People Stores their confidential data on cloud storage and Cloud Service is an Open Service on internet which allows everyone to use Cloud Storage. In this regard security on Cloud is first necessity. This paper aim is to compute matrix multiplication result over malicious cloud. The aim of proposed approach is that we can build a new hash algorithm which can provide a hash value. In proposed approach, we can multiply hash value with matrices for multiplication, result of matrix multiplication with hash value will send to client side. On client side result will be verify after Re transformation, if its correct it will be accepted, otherwise will be aborted.**

Keywords- Hash algorithm, Venomous, Preservation.

I. INTRODUCTION

Cloud computing depends upon sharable/distribute resources to unambiguous and economies of scale, similar to public access. Cloud computing is an technology methodology that facilitate comprehensive approach to shared pools of shared system assets and more advanced-level utilities that will be vastly provisioned with nominal man power, often on the network[1, 2].

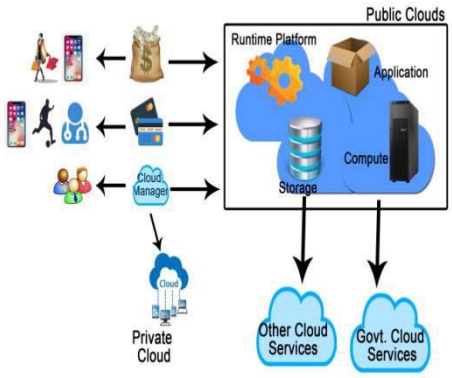


Fig. 1 Cloud computing assistance in Real life Scenario [1]

1.1 Multiple cloud services:

Computing of data on cloud can be done by three ways: infrastructure as a service (providing infrastructure), platform as a service (providing platform) and software as a service (providing Software) [3]. We can represent these services in form of stack, because they fulfill the property of LIFO. Understanding why these services are used and for which purpose for business [4].

1.1.1 Infrastructure as a service

Infrastructure as a service is first service we are discussing, we can pay for Information Technology framework-servers and implicit machines, repository, systems, OS—by a cloud service organizer when required services, then pay for required services[5].

1.1.2 Platform as a service

PaaS specify to cloud computing utilities that helpful as on appeal surroundings for to create, test, applications. PaaS is created to make it comfortable for developer to develop, test and use mobile and desktop apps without investing on or to manage full set up required for development. There is no need to purchase own memory, network. Just use services and pay for that services [6].

1.1.3 Software as a service

SaaS will provide mechanism for delivering on demand applications on the network when required. With Software as a Service, cloud services worker administer the applications and basic framework and control care of applications, i.e. software up gradation and preservation patching. Users can link with the application on the network, mostly on a browser by using mobile and desktop [6].

1.2.1 Categories of cloud developments:

a) **Public**
b) **Private**
c) **Hybrid**

Clouds developments are not similar for every cloud. Provider can developed cloud in their own manner using three ways like public, private and hybrid.

1.2.1 Public cloud

Public clouds can be defined as, hold and managed by a cloud worker, which provide gauge assets i.e. servers and storage on the network Amazon can be called as a public cloud. On public cloud, various types of hardware, software and other necessity infrastructures are hold and managed by some cloud worker people. User can use all services and perform their own account operations using a browser technique [7].

1.2.2 Private cloud

Private cloud can be defined as; means in cloud computing, cloud assets used solely through any organization .private cloud can be substantially situated on the organization location side [7]. Many organizations can pay any other party assistance workers to maintain their private (personal) cloud. Private cloud can be defined as on which the utilities and framework are managed on personal (private) network.

1.2.3 Hybrid (public and private) cloud

Hybrid clouds can be defined as; are one in which there is jointly works as of public and private clouds, joint

With each other by automation which allows information and on demand applications to be accessed by them [7]. By providing access, in which information and on demand applications to be plying common on private and public clouds, hybrid cloud can be defined as useful for businesses greater resilience (flexibility) and many development options. Then optimization achievement is also a concern for researchers due to limited resources hired by client at their end[20].

II. ATTACKS ON CLOUDS

Continuously growth of Cloud usage, attract human to bypass various attentions. It's harmful to use cloud without paying attention to security [8]. Attackers continuously targeted service providers. Vulnerabilities are found in Cloud known as attack. Various types of Vulnerabilities are in Cloud environments.

A. Denial of Service attacks(DOS attacks)

Various type of services provided, may be prevent users from accessing it is known as DOS attack [9]. Cloud itself keeps editing more power to make this attack stronger. Because as it known Cloud is a service provider sometime it exempted some users to provide various types of services.

B. Malware Injection Attack: This attack relies on adding/injecting a service implementation virtual

machine on Cloud scenario. The ambition of this type of attack is to access victim's data on cloud, so the attacker transfer a arrange image and fraud the image to be part of the sufferer cloud environment. After the adverse system/service is added to the cloud environment, user requests will start forwarding to it causing the vulnerable code to execute [10].

C. Brute forcing: In this, the attacker brings up

Virtual Machine and then checks for target in a Zone repeatedly. For VM's generate up in wrong Zone, attacker shut down that Virtual Machine and rerun the process.

D. Authentication and MiTM Attack: As most of the earlier services being offered depends on username/password combination, authentication is assumed to be the frail point in Cloud Security Model. Also if attacker can place themselves between the user and the service provider then the MiTM attacks are also possible [11, 12].

E. Man-in-the Cloud attack:-it is a recent type of attack that focuses on manipulation and extortion user's Cloud synchronization token. Victim is usually hit by malware by sending fraud mails or websites. After that user take control of personal files of victim and useful data. User real synchronization token access by hacker for control purposes [13].

III. PRESERVATION CHALLENGES FOR CLOUD COMPUTING

A. Authorization

Authorization can be defined as a preservation system used to complete approach right or approach levels in contrast of system assets, having computer applications, data, services and application aspects. Authorization can be defined as basically authentication for client identity evidence [14, 15]. In cloud computing scenario user data stored on cloud. So there must be a strong mechanism to provide authorized access control on stored data. After a completing authentication, an authorization technique should provide access rights to requesters.

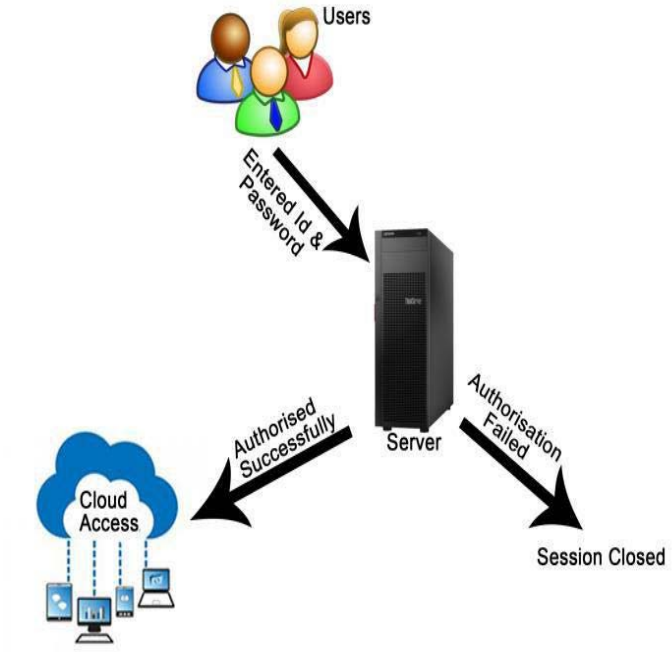


Fig 2. Process of Authorization

B. Confidentiality

It permits authorized users to collect useful and secured data. Some techniques ensure confidentiality and protect data from harmful mediocre [16].

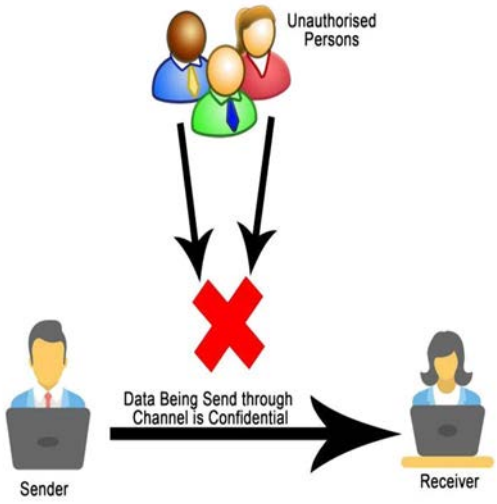


Fig 3. Confidential information

There must be any encryption mechanism on sender side, so that data is protected and send to receiver side. In between path unauthorized individual could not be access your confidential data.

C. Rectitude

Rectitude (Integrity) is the system which will assure digital data is not corrupted (Completed) and will be used or updated by users which will be authorized [16].

Integrity describes preserve the firmness, efficiency and adherence of information on its complete life span.

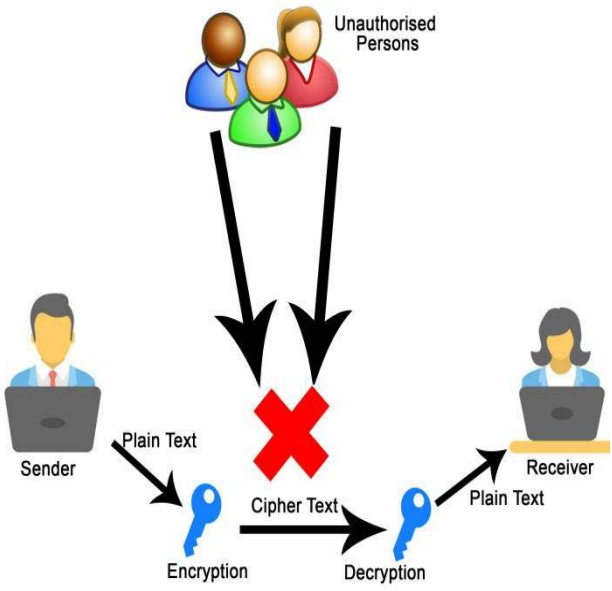


Fig 4. Maintain rectitude of data

D. Translation

Translation technique is necessity technique to keep data secure. In today's life various translation techniques is implementing like QR code, one time password means to send data in encrypted form and on destination side to decrypt it with any suitable algorithm [16].

IV. RELATED WORK

In previous works outsourcing technique was used to improve efficiency of algorithm and to overcome the complexity of algorithm. Author proposed a one way hash algorithm for calculating the result of matrix multiplication. It was one way approach means for encryption and decryption same has value will be used. Concept of invertible matrix was not used like Hill Cipher because in Hill Cipher technique inverse of matrix is used for decryption purpose [12].Author proposed a MCM (Matrix chain multiplication) technique for decreasing complexity of MCM by computing it on Cloud. Complexity of algorithm is decreasing to $O(n^2)$ but original Complexity of MCM is $O(n^3)$.Authors proposed this algorithm on malicious Cloud suspecting that input of computation and output of computation may be leaked and altered. For securing input and output matrices encrypted with some matrices and later it, after receiving result from cloud it will decrypted with invertible matrix for to obtained result. If result received from Cloud is not verified and matched, then result will be aborted [13].Author proposed revised Hill Cipher technique in this technique, Author proposed that in Hill Cipher technique for decrypting result invertible matrix is used, but some time invertible matrix not provided accurate result. So in proposed approach authors used a different key for different matrices [14].Author proposed an outsourcing algorithm to compute characteristics polynomial and eigen value of matrix on Cloud .For securing input and output on malicious cloud, Author proposed to choose random variable from set of values and encrypted original inputs for calculating characteristics polynomial and eigen value on Cloud .Author also proves an algorithm have

accurate for computation, soundness and verifiability[15].

V. PROPOSED PLATFORM

In proposed model one way hash function is generated, which is used for encryption and decryption. From Client side two matrices ($M_{p0} * P1$ and $M_{p1} * p2$) can be send with hash value attached in path to server side. Server can return computation result ($R'_{P0} * p2$). This result will be Retransform into original result ($R^{P0 * p2}$) with calculated hash value before send to client side. If received result is equal to required result, then it will be accepted otherwise result will be aborted on client side. In proposed approach a hash value is attached with matrices for encryption and for mechanism of decryption will be used some other different technique. Modulus value is using with two matrices for encryption so that two input values may not be leaked and output cannot be altered by suspicious cloud. A computation result will be received from Cloud and will be decrypted by client. If it is not same as expected result, then it will be aborted and if it is same as expected result will be verified after Retransformation, and then it will be accepted. Hash value will be computed by proposed hash algorithm.

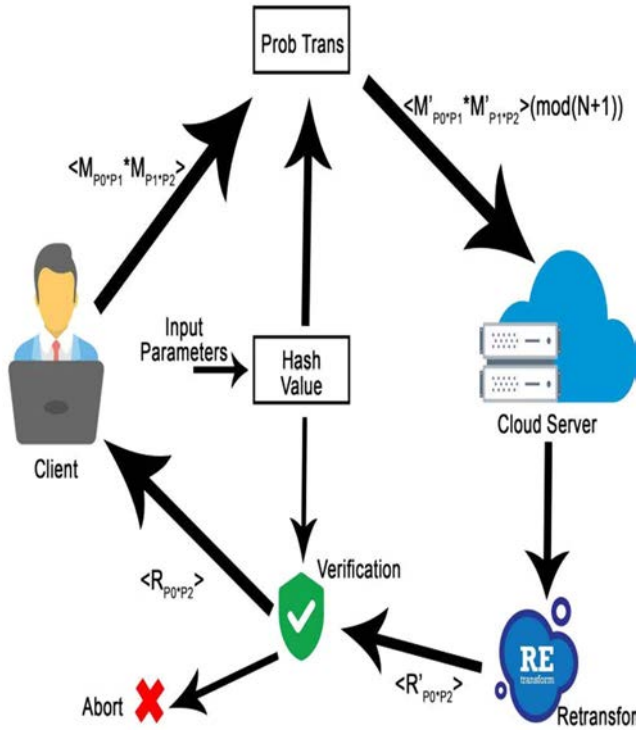


Fig 5.System Model for matrix Multiplication

5.1 Algorithm Framework

Gennaro et al. [17] has presented a formal definition For secure outsourcing algorithm. The outsourcing/deploy algorithm having five strides by the following sequence:

"KeyGen", "ProbTrans", "Compute", "Verify", and "Retransform".

1. ***KeyGen*(1^{λ}):** This work invokes with an input of security criterion λ. Later generates random keys in the form of matrices, which are further used for problem transformation operation.

2. ***ProbTrans*** **(ϕ, k):** On the input of security key matrices *k*. The *ProbTrans* (ϕ, k) operation perform the encryption, the original problem $\phi\ (M_{i-1}, M_i, k)$ with the key *k* generates the translate problem $\phi k\ (M'_{i-1}, M'_i, k)$

3. ***Compute*** **$(\phi k\ (M'_{i-1}, M'_i, k))$:** The server performs the computation and produces the encrypted result.

4. ***Retransform*** **(R'_i, k):** when cloud server completed the execution of problem, it returns the decision to client. The client with the help of keys (k) retransform the result and find $R_i = \phi\ (M_{i-1}, Mi, k)$.

5. ***Verify*** **(Ri, k):** Finally, the client with the help of security keys verifies the correctness of result *Ri*.

In proposed algorithm it is suspected that cloud server can be produce incorrect result, because server can be malicious.

5.2 Design Model

(a) Hash Value (Key): In proposed model one way hash function is generated, which is used for encryption and decryption [18][19].

$$H(V) = V \bmod (N+1)$$

H (V) = Hash value in matrix form with modular value

V= n * n matrix

N= modular value

We can take V as matrix which will be used to compute Hash value. We can take value of R in form of n* n matrix as input and value of N also as an input parameter.

(b) ProbTrans : Apply secure hash value to matrices ($M_{p0 * P1}$ and $M_{p1 * p2}$), which will be obtained from hash function.

(c) Compute: Server will compute result of two matrices and hash value applied to matrices.

$$R'_{p0 * p2} = H(V) * M_{P0 * P1} * M_{P1 * P2}$$

$$= V * (M_{P0 * P1} * M_{P1 * P2}) \pmod{(N+1)}$$

$$= M'_{p0 * p2} \pmod{(N+1)}$$

Where M' P0 * P1 * M' P1 * P2=

V * M P0 * P1 * M P1 * P2 and

$$M'_{p0 * p2} = V * M_{P0 * P1} * M_{P1 * P2}$$

$$R'_{p0 * p2} = M'_{p0 * p2} \pmod{(N+1)}$$

(d) ***Retransform:*** when cloud server completed the computation of results, it returns the result to client. Then received result will be transform into original required result by:

$$R_{p0*P2} \bmod (N+1) = V^{-1} R'_{p0 * p2}$$

(e) ***Verify:*** Finally, the client with the help of security keys verifies the correctness of result $\mathbf{R_{p0 * p2}}$.

For proving this result we have:

$$R_{p0*P2} \bmod (N+1) = V^{-1} \; R'_{p0 * p2}$$

$$R_{p0*P2} \bmod (N+1) = V^{-1} \; V * (M_{P0 * P1} * M_{P1 * P2}) \bmod (N+1)$$

$$\text{Here } R'_{p0 * p2} = M'_{p0 * p2} \pmod{(N+1)},$$

$$M'_{p0 * p2} = V * M_{P0 * P1} * M_{P1 * P2}$$

$$\text{and } VV^{-1} = 1$$

$$\text{Hence } R_{p0*p2} = M_{P0 * P1} * M_{P1 * P2}$$

VI. CONCLUSION

This paper discussed about these two issues. Today's many organizations; government sectors are suffering with security issues and searching various techniques to tackle with Security concern. Security era in cloud computing is very wide. In proposed model computation of matrix multiplication is done with the help of hash value. Hash value is computed form matrix and modulo parameter is attached with matrix. But during computation of results, we are multiplying our two matrices with modulo and then multiplying value of n*n matrix. In proposed approach matrix multiplication outsourcing algorithm is performing on serve side. Server is malicious, for to secure our result, we have applied hash algorithm on matrices to get correct result from server. The proposed approach is able to meet the design goals of truth, security, efficiency and verifiability. In future we can implement proposed approach on more than two matrices. In future we can compute this algorithm on standalone desktop and also on server for to match complexity difference and performance gain between results of Server side and standalone desktop.

VII. REFRENCES

[1] Nasrin Delil and Ahmed Kayed, **"**Preserving data in Cloud Computing**"**, IJCSI, Vol. 12,Issue 2,March 2015.

[2] Wolfgang Lehner and Kai-uwe sattler,**"** Database as a Service**"**, ICDE, 2010.

[3] Michael Armbrust, Armando Fox and Rean Griffith,**"**A view on Cloud Computing**"**, ACM, Vol. 53 No. 4, PP. 50-58, April 2010.

[4] Jian Shen, Tianqi zhou, Xiaofeng Chang, **"**Anonymous and Traceable Group data sharing in Cloud Computing**"**,IEEE, Vol. 13, Isssue 4, 2018.

[5] Yi, Xun, Russell Paulet, and Elisa Bertino. **"***Homomorphic encryption and applications***"**, Vol. 3. Heidelberg: Springer, 2014.

[6] Wang, Liangmin, Zhendong Yang, and Xiangmei Song. "SHAMC: A Secure and highly available database system in multi-cloud environment." *Future Generation Computer Systems* (2017).

[7] Wenying Zeng, Yuelong Zhao and Junwei Zeng **"**Cloud Service and Service Selection Algorithm Research**"**, ACM, 2009.

[8] Anandaraj, S. P., and Mohammed Kemal. "Research opportunities and challenges of security concerns associated with big data in cloud computing." In *I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC), 2017 International Conference on*, pp. 746-751. IEEE, 2017.

[9] Fox Robert**"**Library in the clouds OCLC Systems & Services**"**, IDLP, Vol. 25 , Issue 3, 2009.

[10] Dr. Firas A. Abdul Atif and Maan **"**Cloud Security issues and challenges important points to move towards cloud storage**"**, IJSR, August 2017.

[11] B. Acharya, G. S. Rath, and S. K. Patra, **"**Novel Modified Hill Cipher Algorithm,**"** *Int. Conf. Emerg. Technol. Appl. Eng. Technol. Sci.*, pp. 126–130, 2008.

[12] X. Hu and C. Tang, **"**Secure outsourced computation of the characteristic polynomial and eigenvalues of matrix,**"** pp. 4–9, 2015.

[13] L. Zhou, Y. Zhu, and K. R. Choo, **"**Efficiently and securely harnessing cloud to solve linear regression,**"** *Futur. Gener. Comput. Syst.*, 2017.

[14] C. Pramkaew and S. Ngamsuriyaroj, **"**Journal of Information Security and Applications Lightweight scheme of secure outsourcing SVD of a large matrix on cloud,**"** *J. Inf. Secur. Appl.*, vol. 41, pp. 92–102, 2018.

[15] U. Gupta, M. S. Saluja, and M. T. Tiwari, **"**Enhancement of Cloud Security and removal of anti-patterns using multilevel encryption algorithms,**"** 2018.

[16] C. Kaleeswari, P. Maheswari, K. Kuppusamy, and M. Jeyabalu, **"**A Brief Review on Cloud Security Scenarios,**"** vol. 4, p. 7, 2018.

[17] R. Gennaro, C. Gentry and B. Parno, Non-nteractive verifiable computing: Outsourcing computation to un trusted workers, *Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics) LNCS* **6223** (2010), 465–482.

[18] M. Farajallah, M. Abu Taha, and R. Tahboub, **"**A Practical One Way Hash Algorithm based on Matrix Multiplication,**"** *Int. J. Comput. Appl.*, vol. 23, no. 2, pp. 34–38, 2011.

[19] M. Kumar and M. Vardhan, **"**Data confidentiality and Integrity preserving outsourcing algorithm for matrix chain multiplication over malicious cloud server,**"** in *Journal of Intelligent and Fuzzy Systems*, 2018.

[20] Lal G, Goel T, Tanwar V, Tiwari R. Performance Tuning Approach for Cloud Environment. In The International Symposium on Intelligent Systems Technologies and Applications 2016 Sep 21 (pp. 317-326). Springer.

Energy-aware Virtual Machine Selection and Allocation Strategies in Cloud Data Centers

Harvinder Singh
Department of Virtualization
School of Computer Science
University of Petroleum and Energy Studies
Dehradun, Uttarakhand
Email: hsingh@ddn.upes.ac.in

Sanjay Tyagi
Department of Computer Science and Applications
Kurukshetra University
Kurukshetra, Haryana
Email: tyagikuk@gmail.com

Pardeep Kumar
Department of Computer Science and Applications
Kurukshetra University
Kurukshetra, Haryana
Email: mittalkuk@gmail.com

***Abstract*—These days, information technologies are expanding exponentially so the need for high-speed processors and huge storage space are developing quickly. As a result of increasing requests, more resources are required to satisfy the client's necessities. Thus, in a cloud environment, the large number of resources consumes a lot of energy during their operation, which is turned into a key issue nowadays and demands a critical discussion in the present scenario. This research paper investigates and explores the literature on the assignment of virtual machines to hosts in a data center according to variable workload requests of different cloud consumers application executing on the virtual machines. The choice of ideal virtual machines and their placement on host prompts to limit the energy utilization. This paper proposes an algorithm for improving virtual machine selection and allocation strategies in cloud data centers. The proposed algorithm is then compared with existing algorithms on the basis of performance metrics like energy consumption and VM migrations using different threshold values. As a result, the proposed algorithm emerged to be the optimized one in enhancing the use of cloud resources by lessening the energy utilization of datacenter.**

I. Introduction

Cloud environment comprises of various data centers that contain millions of physical machines including a substantial number of cloud resources like processing elements, memory and network assets for various cloud consumers applications. Cloud computing is the advancement of different computing techniques like a grid, distributed and parallel computing or characterized as the business usage of these ideas. Cloud computing is a rising technology with an extensive accumulation of heterogeneous self-ruling assets with adaptable computing frameworks. This computing has altered the IT industry by empowering on-request allocation of cloud assets on a compensation as-you-go basis.

Cloud computing has brought about the foundation of cloud data centers at vast scale containing a large number of computing nodes. Moreover, data centers in the cloud expend colossal amount of energy in the form of electricity, bringing about high working expenses and carbon dioxide emanations to the surroundings. Moreover, carbon dioxide discharges from the IT business is essentially adding to the greenhouse effect [1] [2]. Some remedial steps are needed to be taken in order to limit the energy utilization in data centers by utilizing effective and efficient resource management techniques [3] [4].

To reduce the high energy consumption, it is important to wipe out wasteful aspects, like manner in which power is conveyed to processing assets, and in the manner in which these assets are used to serve cloud consumers applications. This should be possible by enhancing both the physical framework of the cloud data center and the asset distribution and administration techniques.

II. Motivation

The VM selection and allocation strategies are based on aligning VMs to PMs depending on their characteristic requirements. In order to maintain a balance between energy consumption and overloading of PM. it is recommended to suggest different techniques for reducing the PM underutilization and the PM overutilization for cloud data centers.

A. Literature Review

In this section, the background of the VM selection and placement strategies are discussed, that further results in the inception of the proposed modified threshold based host overload detection and minimum migration time algorithm. A lot of literature is focussed on figuring different successful strategies of VM's migration for reducing the power utilization at overseeing host servers without suggesting that under SLA infringement, hence proposing three of the arrangements for the same.

The first arrangement is named as Minimization of Migration(MM) which additionally works with a couple of thresholds (most extreme and minimum)for dropping the amount of relocation of VM among PMs [5]. In this arrangement, if at the server end, the CPU utilization is underneath to that of the minimum limit, at that point all the VM's for this situation will relocate to the other server and the sender server will then be shut off. Anyway, when the most extreme limit is surpassed by the CPU utilization of the server, for this situation just a few of the VM's will be relocated. In both these cases, the relocation of VM's to the server is called as Highest Potential Growth (HPG) strategy, that inturns set the PMs to get new VM's on basis of most minimal CPU usage rate [6].

The other strategy named as Choice and Random (RC) Policy which prompts relocation of VM's with that to a consistently disseminated random variable. The result demonstrated us that the Dynamic VM shutdowns and relocations out of idle

978-1-5386-6026-3/18/$31©2018 IEEE

PMs bring an enormous measure of saving energy at the data center [7]. Also, it expresses that on use of these strategies, one can spare energy up to 83% at the point when contrasted with Non-Energy Aware Strategy which in turn prompts 66% sparing at the point when additionally contrasted with Dynamic Voltage and Frequency Scaling procedure and at last more than 23% of sparing when contrasted with Simple Threshold Technique [8].

Additionally, a VM scheduler works in two stages. The beginning prompts the choice of VMs by checking and controlling a dynamic threshold picked for CPU usage. The subsequent stage is about Best Fit Decreasing strategy which selects a location for distribution of VM's after relocation [9]. The utilization of dynamic threshold with dynamic tasks at hand attempts to acquire a drop in utilization of energy and further to limit the violation events of SLA in the cloud [10]. The results additionally demonstrate that when a dynamic threshold is used, there is a decrement in SLA infringement occurrence by 25% when contrasted with a single threshold [11]. Attempting to diminish the quantity of CO2 transmitted out in the environment is one of the essential requirement of VM scheduler design [12]. The performance of the algorithm used for mapping the task submitted by the users, to the requested resource must be effective, thereby helping in reducing the number of VMs relocations and SLA violations [13].

As saving energy is the prime objective of the VM scheduler, which can be achieved by choosing best PM for mapping of VMs in order to execute the task requested by the cloud consumers in HPC cloud environment [14]. Ye et al. [15] proposed technique in could lead to 35% fall in the absolute utilization of energy contrasted with the best in class distribution heuristics.

During the literature survey, it is noted that grading traditional VM placement algorithms [16] [17] [18] [19] or expressing the best one out of the many isn't an appropriate proposal in light of the fact that each one of other VM Placement method has some particular target, migration method, major resources, and credible parameters [20]. Despite the fact that these parameters may appear to be fine from external view, there may have a few or the other sort of tradeoffs when profoundly overviewed [21]. Inferable from the variations in workload and changing the structure of consumers applications, there is an obligation to persistently reform the traditional VM placement techniques.

B. Challenges

Specifically, the accompanying VM-PM mapping issues are examined from the literature:

1) What are the different ways to determine the most effective method to characterize workload-free QoS necessities.
2) How to determine the stage at which VMs can be relocated from over usage servers to stay away from deterioration in performance while fulfilling the characterized QoS requirements.
3) How to determine the stage at which VMs can be relocated from under usage servers to enhance usage of assets and limit utilization of energy while fulfilling the characterized QoS requirements.
4) Shifting of which VM should be done.
5) Where to move the VMs chose for relocation.
6) How to determine the stage at which physical machines should be turned on/off.
7) How to design the most effective techniques for VM solidification in a cloud environment.

C. Problem Statement

The main role of the cloud environment is that its customer can use the assets to have monetary advantages. An asset mapping process is required to keep away from under usage or over usage of assets which may influence the performance of cloud services.

This research paper proposes techniques to handle the challenges in connection to energy proficient VM to PM mapping in cloud environment subjected to QoS limitations.

D. Objective of the Research Paper

To manage the difficulties related to the research issues discussed in the challenges section, the present research paper has the following goals:

1) To study and explore various existing host overload detection algorithms.
2) To propose a threshold based host overload detection algorithm along with minimum migration time algorithm.
3) To compare and validate the proposed algorithm by using cloud simulation tool.

III. Proposed System Model

In a cloud environment, services are given to cloud consumers continuously on request. The number of users accessing the cloud environment is always more than that was using it on the previous day. Cloud consumers will develop applications to be run on cloud infrastructure, cloud providers will provide and manage infrastructure for running the jobs submitted by the consumers. Moreover, the requests of cloud consumers for accessing cloud services are increasing exponentially these days. Further, for servicing these large number of requests huge cloud infrastructure like servers, storage space, networks etc are required.

In this paper, the effective procedure is proposed to adjust the load on cloud VMs and PMs, with a goal that the VMs and PMs don't get crash and they can hold on long which in turn results in less energy consumption. Utilization of energy isn't just dictated by equipment proficiency yet in addition to the asset management framework deployed on the cloud infrastructure and the productivity of consumers jobs running in the framework. Later on, the proposed algorithm is implemented on cloudsim simulator and then compared with the existing VM consolidation algorithm to validate the outcomes.

A. Method Description

The method proposed in the present paper will help to measure the performance of the cloud system when the cloud consumers requests are scaled up substantially. The estimations

empowered the perception of VMs conduct and in addition the elaboration of the focal thought of a scaling strategy for VMs that run consumers applications.

A noteworthy challenge of building up a scheduler is consolidating the parameters like utilization of energy and performance of VMs. In the datacenter servers, to distribute assets and execute the cloud applications submitted by the cloud consumers, the proposed algorithm must consider the utilization of energy by the cloud resources. In this manner, the proposed algorithm works on energy consumption as well as the performance of the allocated cloud resources like VMs and PMs to the requested cloud applications.

B. Research Methodology

To develop the scheduling technique, some steps are needed to be followed. These steps are the choice of the tools, the information gathering, and testing and the assessment of the utilization of the energy. In this section, all the general building procedure of the proposed algorithm is portrayed. Moreover, the points of interest of the proposed algorithm are clarified in the following subsections.

1) Experiments to collect background data: At first, tests were performed on information gathered, related to utilization of energy by different VMs that are executing various cloud applications. So this perception could deliver characteristics which the proposed algorithm would consider. The gathered information was utilized as a basis to make the assignment approach followed in the proposed algorithm considering the sort of VM to be stunned and the kind of applications that a VM will execute.

Along these lines, it was conceivable to limit under and over-utilization of assets, looking to address the trade-off between VMs performance and energy utilization applications. In any case, the best load was built up for the VMs and after that rehashed this charge to alternate VMs, notwithstanding the count of VMs resources like processors and memory. This measure was additionally done so as to get a performance proportion (runtime) per energy cost (total Watts).

2) Measuring energy consumption: The authentication of the proposed algorithm can be determined by the analytical computation of expended assets, concerning both inhabitance level and utilization of energy results in the conditions and points of confinement of VMs dissemination on the server. Since every application has an alternate behavior as per the computational asset utilized, such data was the parameter for the underlying improvement of the proposed algorithm. In this way, checking and recording the contrasts in cloud applications is the most critical aspect for deciding which VM to be relocated to which PM.

3) Choosing a measuring method: The proposed method is centered around the relocation of VMs. For this, it is required to identify and recognize which VM has a more outstanding burden, which VM has less load and which has no assignment. In view of these three conditions, the relocation of VM should be done.

4) Mathematical model of proposed System: The proposed approach centers around the movement of VMs to PM having a high outstanding load, which PM has low load and which has no task for execution. For migration of VMs, various parameters are computed quantitatively by using the following formulae [19].

$$ram\_utilization = \frac{total\_requested\_ram\, from\, vm}{total\_ram\, of\, host} \quad (1)$$

$$mips\_utilization = \frac{total\_requested\_mips\, from\, vm}{total\_mips\, of\, host} \quad (2)$$

$$bw\_utilization = \frac{total\_requested\_bw\, from\, vm}{total\_bw\, of\, host} \quad (3)$$

$$total\_utilization = \frac{1}{3} * (ram\_utilization * bw\_utilization * mips\_utilization) \quad (4)$$

$$ram\_utilization = \frac{\sum_{i=0}^{n} ram\, of\, vm_i}{total\_ram\, of\, host} \quad (5)$$

$$mips\_utilization = \frac{\sum_{i=0}^{n} mips\, of\, vm_i}{total\_mips\, of\, host} \quad (6)$$

$$bw\_utilization = \frac{\sum_{i=0}^{n} bw\, of\, vm_i}{total\_bw\, of\, host} \quad (7)$$

where, ram_utilization = Memory utilization of host, mips_utilization = CPU utilization of host, bw_utilization = Bandwidth utilization of host

5) Proposed Algorithms: At the point when VMs are over-burden, because of the exhaustion of a physical servers assets that have the VMs, energy utilization is expanded. The technique selected for mapping of VMs to PMs directly affects the utilization of energy of the data center. So it must guarantee that the VMs are mapped on vast capacity PMs in order to maintain throughout a balanced performance. The algorithm for picking over-utilized host is given as Algorithm 1.

$$occupied\_resource\_weight\_ratio = \frac{\sum_{i=0}^{n} resource\_weight\ of\ vm_i}{\sum_{i=0}^{m} available\_resource\_weight\, of\, running\, host} \quad (8)$$

Occupied resource weight ratio determines from equation 8 identifies which PM is over-burden or under-burden or has no task to execute. Furthermore, if PM has no task mapped to it, then it should be switched off. And if the PM is underloaded then relocate it to another PM having less number of tasks for execution and shut off that PM until the new tasks arrived. In contrast to this, if the PM is overloaded then relocate VMs from that PMs to other running PMs. The algorithm for selecting VM for migration from a host and the algorithm for finding a host for VM allocation is given as Algorithm 2

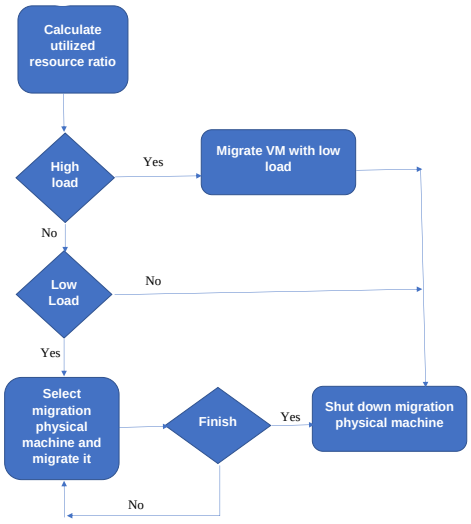


Fig. 1. Flowchart for the proposed approach

and Algorithm 3 respectively. The flowchart for the proposed approach is shown in Figure 1.

PMs are considered as overburden if the gross occupied resource weight ratio is greater than the upper threshold value, and if it is less than the upper threshold value then that PMs are considered as underburden. This is the methodology adopted for optimization of utilization of energy by VMs.

Algorithm 1 Picking Over-Utilized Host

1: threshold=user_defined_threshold
2: Initially overloaded_pm_list doesn't contain any host.
3: Calculate ram_utilization, mips_utilization, bw_utilization and total_utilization of host
4: **for** each host h **do**
5: Calculate $ram\_utilization \leftarrow \frac{total\_requested\_ram}{total\_ram\_of\_host}$
6: Calculate $mips\_utilization \leftarrow \frac{total\_requested\_mips}{total\_mips\_of\_host}$
7: Calculate $bw\_utilization \leftarrow \frac{total\_requested\_bw}{total\_mips\_of\_bw}$
8: Calculate $total\_resource\_utilization \leftarrow \frac{(ram\_utilization * mips\_utilization * bw\_utilization)}{3}$
9: **if** $(total\_resource\_utilization > threshold)$ **then**
10: Add h in overloaded_host_list

IV. Implementation and Results

Utilization of energy by the assets of cloud environment expands exponentially and turn into a key issue in the cloud. Data centers in cloud environment devour immense quantity of energy and furthermore emanate carbon dioxide in nature. For optimization of energy utilization, energy effective asset management is required. The present paper proposed a novel approach to deal with the energy use of VMs.

A. Experimental Setup

The empirical examination of the proposed approach has been done and outcomes are acquired by programming this approach inside the Cloud Sim simulation tool. Each VM has its own particular qualities and utilization highlights relying on the use of assets and in this way produce divergent amounts of carbon dioxide. The aggregate of carbon impression of the data center is with respect to the utilization of energy by each host.

Algorithm 2 Selecting Virtual Machine for Migration from Host

1: utilization[ram] define utilization of memory, utilization[mips] define utilization of mips and utilization[bw] define utilization of bandwidth
2: $vm\_migration\_list \leftarrow null$
3: $maximum \leftarrow 0$
4: **if** $(mips\_utilization < ram\_utilization)$ **then**
5: **if** $(bw\_utilization < ram\_utilization)$ **then**
6: **for** each Virtual Machine vm in overloaded_host **do**
7: **if** $(maximum < utilization[ram])$ **then**
8: $maximum \leftarrow utilization[ram]$
9: $vm \leftarrow vm_i$
10: Add vm in vm_migration_list
11: **if** $(bw\_utilization < mips\_utilization)$ **then**
12: **for** each Virtual Machine vm in overloaded_host **do**
13: **if** $(utilization[mips] < maximum)$ **then**
14: $maximum \leftarrow utilization[mips]$
15: $vm \leftarrow vm_i$
16: Add vm in vm_migration_list
17: **if** $(bw\_utilization < mips\_utilization)$ **then**
18: **for** each Virtual Machine vm in overloaded_host **do**
19: **if** $(utilization[mips] < maximum)$ **then**
20: $maximum \leftarrow utilization[mips]$
21: $vm \leftarrow vm_i$
22: Add vm in vm_migration_list
23: **for** $eachVirtualMachinevm_iinoverloaded\_PM$ **do**
24: **if** $(utilization[bw] < maximum)$ **then**
25: $maximum \leftarrow utilization[bw]$
26: $vm \leftarrow vm_i$
27: Add vm in vm_migration_list

Algorithm 3 Finding Host for Virtual Machine Allocation

1: Algorithm returns true if suitable host found
2: $selected\_pm \leftarrow null$
3: check each host p from host_list
4: **if** enough resources on p for vm **then**
5: allocate vm on host p

B. Results Analysis & Discussion

The mapping of a large number of VMs on the same PM encourages in solidifying the mission and shutting off other PMs which will chops down the energy utilization level considerably. The proposed algorithm is tested and compared by considering energy consumption and a number of VM migrations as performance metrics.

1) Energy Consumption: The comparative analysis of existing and proposed algorithm based on energy consumption (in kWh) vs the number of cloudlets(n) is shown by the graphical representation in Figure 2 corresponding to the value obtained in Table 1 after empirical examination.

TABLE I. COMPARATIVE ANALYSIS OF EXISTING AND PROPOSED ALGORITHM BASED ON ENERGY CONSUMPTION

Number of Cloudlets	Energy Consumption using Existing Method	Energy Consumption using Proposed Method
100	54.64	14.69
150	56.76	19.13
200	59.2	26.05
250	59.7	31.37
300	60.2	37.27

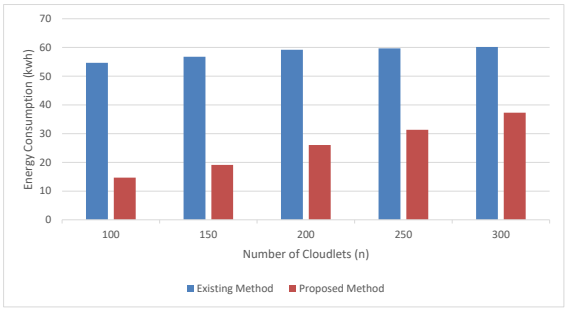


Fig. 2. Comparative Analysis of Existing and Proposed Algorithm based on Energy Consumption

TABLE II. COMPARATIVE ANALYSIS OF PROPOSED ALGORITHM WITH DIFFERENT THRESHOLD VALUES BASED ON NUMBER OF VM MIGRATIONS

Number of Cloudlets	Number of Migrations using Proposed Method with Threshold = 0.6	Number of Migrations using Proposed Method with Threshold = 0.9
500	61	1230
600	59	1051
700	51	341
800	44	246
900	43	210

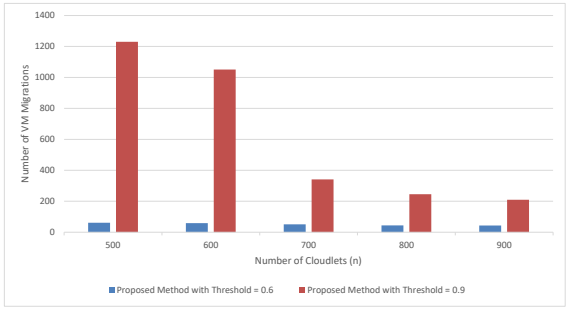


Fig. 3. Comparative Analysis of Proposed Algorithm with different threshold values based on Number of VM Migrations

Discussing the existing method, the graph clearly conveys that the energy consumption is almost constant and very high with respect to small increase in the number of cloudlets. The energy consumption for 100 cloudlets is 54.64 kwh while for 300 is 60.2 kwh i.e. ranging between 54 to 61 kwh for 100 to 300 cloudlets with almost increase of 1 kwh for each 50 increase in the number of cloudlets. On the other hand, the proposed method is very conservative in terms of energy consumption, as it consumes only 14.69 kwh for the 100 number of cloudlets and with reaching almost a bit high from the half of previous method's maximum i.e. almost 38 kwh for 300 cloudlets in present, which shows the energy consumption efficiency of the proposed method.

Hence, when compared on the basis of energy consumption, the proposed method is highly optimized in contrast to the existing method. From the graphical analysis following observations has been extracted.

1) Highly dependent on the number of cloudlets (not almost constant like the existing method).
2) Low for less number of cloudlets and grows in a proper way for increasing number of cloudlets.
3) Good for especially less number of cloudlets as energy consumption is very low from the start rather than the existing method where consumption is very high for even small number of cloudlets.

2) Number of VM Migrations: The comparative analysis of proposed algorithm based under two threshold values; 0.6 and 0.9, on the number of VM migrations vs the number of cloudlets(n) is shown by the graphical representation in Figure 3 corresponding to the value obtained in Table 2 after empirical examination.

The results depict that the value of threshold plays a key role. Discussing for the threshold 0.6, the number of VM migrations remain almost constant; ranging between 43 to 61; for the increasing number of cloudlets; between 500 to 900. On increasing the threshold to 0.9 for the same range of cloudlets, the graph shows a drastic increase in the number of migrations which reaches almost 1230 for 500 number of clouds while on increasing cloudlets from 600 to 700 , the graph shows a huge surge from almost 1051 to 341 (although value for number of VM migrations still more than what was there for threshold = 0.6 i.e. 51 for 700 cloudlets). On further increasing the number cloudlets from 700 to 900, the number of VM migrations decreasing but still more than what was for threshold 0.6. It shows that increasing the threshold on machine above the 0.6 thresholds will result in some drastic increase in the number of VM migrations especially for less number of cloudlets. Thus, it is recommended to use the proposed algorithm with high threshold values to accommodate huge load on VMs in a cloud environment.

V. CONCLUSION

Cloud computing innovation is required to develop and give vast services and computational energy to end clients. For this, energy efficiency is more critical for virtualized data centers because of more power utilization, higher running cost and a large amount of CO2 discharge to the environment. VM placement in data centers of the cloud environment has been a dynamic zone of research in the last couple of years. This research paper proposed an algorithm for VM selection and placement in cloud data centers. The goal of this approach is the minimization of energy consumption, maximization of resource utilization, enhancement of load balancing improvement of QoS and to achieve green cloud computing. In the latter part of the paper, the proposed algorithm emerges out to be optimum in terms of energy utilization and a number of VM migrations as compared to the existing algorithms.

Furthermore, a more profound examination of data center network architecture and topology, resource properties, and qualities may bring forth different variations as a future scope and can give inventive ways to deal with issues like energy consumption and VM migration.

REFERENCES

[1] E. Oró, V. Depoorter, A. Garcia, and J. Salom, "Energy efficiency and renewable energy integration in data centres. Strategies and modelling review," *Renewable and Sustainable Energy Reviews*, vol. 42, pp. 429–445, 2015. [Online]. Available: http://dx.doi.org/10.1016/j.rser.2014.10.035

[2] T. Baker, B. Al-Dawsari, H. Tawfik, D. Reid, and Y. Ngoko, "GreeDi: An energy efficient routing algorithm for big data on cloud," *Ad Hoc Networks*, vol. 35, pp. 83–96, 2015. [Online]. Available: http://dx.doi.org/10.1016/j.adhoc.2015.06.008

[3] M. Mishra and A. Das, "Dynamic resource management using virtual machine migrations," *IEEE Communications Magazine*, vol. 50, no. September, pp. 34–40, 2012. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6295709

[4] L. Mashayekhy, M. M. Nejad, and D. Grosu, "Physical Machine Resource Management in Clouds: A Mechanism Design Approach," *IEEE Transactions on Cloud Computing*, vol. 3, no. 3, pp. 247–260, 2015.

[5] A. Choudhary, S. Rana, and K. J. Matahai, "A Critical Analysis of Energy Efficient Virtual Machine Placement Techniques and its Optimization in a Cloud Computing Environment," *Physics Procedia*, vol. 78, no. December 2015, pp. 132–138, 2016. [Online]. Available: http://dx.doi.org/10.1016/j.procs.2016.02.022

[6] W. Tian, M. He, W. Guo, W. Huang, X. Shi, M. Shang, A. N. Toosi, and R. Buyya, "On minimizing total energy consumption in the scheduling of virtual machine reservations," *Journal of Network and Computer Applications*, vol. 113, pp. 64–74, 2018. [Online]. Available: https://doi.org/10.1016/j.jnca.2018.03.033

[7] S. Vakilinia, M. M. Ali, and D. Qiu, "Modeling of the resource allocation in cloud computing centers," *Computer Networks*, vol. 91, pp. 453–470, 2015. [Online]. Available: http://dx.doi.org/10.1016/j.comnet.2015.08.030

[8] N. Moganarangan, R. G. Babukarthik, S. Bhuvaneswari, M. S. Saleem Basha, and P. Dhavachelvan, "A novel algorithm for reducing energy-consumption in cloud computing environment: Web service computing approach," *Journal of King Saud University - Computer and Information Sciences*, vol. 28, no. 1, pp. 55–67, 2016. [Online]. Available: http://dx.doi.org/10.1016/j.jksuci.2014.04.007

[9] H. Duan, C. Chen, G. Min, and Y. Wu, "Energy-aware scheduling of virtual machines in heterogeneous cloud computing systems," *Future Generation Computer Systems*, vol. 74, pp. 142–150, 2017.

[10] D. Panagiotou, E. Oikonomou, and A. Rouskas, "Energy-efficient virtual machine provisioning mechanism in cloud computing environments," *Proceedings of the 19th Panhellenic Conference on Informatics - PCI '15*, pp. 197–202, 2015. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2801948.2801989

[11] E. Oikonomou, D. Panagiotou, and A. Rouskas, "Energy-aware Management of Virtual Machines in Cloud Data Centers," pp. 2–7, 2015.

[12] M. Masdari, S. S. Nabavi, and V. Ahmadi, "An overview of virtual machine placement schemes in cloud computing," *Journal of Network and Computer Applications*, vol. 66, pp. 106–127, 2016. [Online]. Available: http://dx.doi.org/10.1016/j.jnca.2016.01.011

[13] H. Shen and L. Chen, "CompVM: A Complementary VM Allocation Mechanism for Cloud Systems," *IEEE/ACM Transactions on Networking*, pp. 1–14, 2018.

[14] I. Rodero, J. Jaramillo, A. Quiroz, M. Parashar, F. Guim, and S. Poole, "Energy-efficient application-aware online provisioning for virtualized clouds and data centers," *2010 International Conference on Green Computing, Green Comp 2010*, pp. 31–45, 2010.

[15] X. Ye, Y. Yin, and L. Lan, "Energy-efficient many-objective virtual machine placement optimization in a cloud computing environment," *IEEE Access*, vol. 5, no. c, pp. 16 006–16 020, 2017.

[16] H. Wang and H. Tianfield, "Energy-Aware Dynamic Virtual Machine Consolidation for Cloud Datacenters," *IEEE Access*, vol. 6, no. c, pp. 15 259–15 273, 2018.

[17] A. Satpathy, S. K. Addya, A. K. Turuk, B. Majhi, and G. Sahoo, "Crow search based virtual machine placement strategy in cloud data centers with live migration," *Computers and Electrical Engineering*, vol. 0, pp. 1–17, 2017. [Online]. Available: https://doi.org/10.1016/j.compeleceng.2017.12.032

[18] B. Gohil, "A Comparative Analysis of Virtual Machine Placement Techniques in the Cloud Environment," vol. 156, no. 14, pp. 12–18, 2016.

[19] M. C. Silva Filho, C. C. Monteiro, P. R. Inácio, and M. M. Freire, "Approaches for optimizing virtual machine placement and migration in cloud environments: A survey," *Journal of Parallel and Distributed Computing*, vol. 111, pp. 222–250, 2018. [Online]. Available: http://dx.doi.org/10.1016/j.jpdc.2017.08.010

[20] A. Vafamehr and M. E. Khodayar, "Energy-aware cloud computing," *Electricity Journal*, vol. 31, no. 2, pp. 40–49, 2018. [Online]. Available: https://doi.org/10.1016/j.tej.2018.01.009

[21] A. Mosa and R. Sakellariou, "Virtual Machine Consolidation for Cloud Data Centers Using Parameter-Based Adaptive Allocation," 2017.

Proceedings of the SMART–2018, IEEE Conference ID: 44078
2018 International Conference on System Modeling & Advancement in Research Trends, 23rd–24th November, 2018
College of Computing Sciences & Information Technology, Teerthanker Mahaveer University, Moradabad, India

Facial Recognition-Application and Future: A Review

[1]Vishwani Sati, [2]Dhruv Garg, [3]Tanupriya Choudhury, [4]Archit Aggarwal
[1,2,4]Amity University, Uttar Pradesh, Noida, India [3]University of Petroleum & Energy Studies, Dehradun, Uttarakhand, India
[1]vishwani.sati@gmail.com,[2]dhruv98garg@gmail.com,[3]tanupriya1986@gmail.com, [4]archit.aggarwal1508@gmail.com

***Abstract*—The whole program comprises of four useful pieces, to be specific 'LoadImages', 'ConstructEigenfaces', 'ClassifyNewface', and 'undoUpdateEigenfaces'. There is likewise a "fundamental" capacity, which calls 'ConstructEigenfaces' and 'ClassifyNewface' capacities to finish the face acknowledgment undertaking. Facial biometrics features depict the ability to be inculcated with any modern camera. The technology has also been used for authentication on smart devices along with many other daily applications like Banking apps, payment apps and simply logical access control. This paper aims to present a detailed overview and review of this technology and the potential future outlines for the technology.**

Keywords: Artificial Intelligence, Machine Learning, Facial Recognition

I. Introduction

The face expects a crucial part in our social intercourse in passing on character and feeling. The human ability to see faces is fundamental. The human capacity to see faces is major. We can see a giant number of countenances learnt all through our lifetime and see trademark appearances at first paying little regard to following quite a while of package. The most remote point is sensible, paying little respect to wide degrees of progress in the visual lift in setting of graph conditions, verbalization, making, and redirections, for example, glasses or changes in hairdo.

Face accreditation is seeing the opportunity to be when in doubt one of the considerable outlined perspectives in PC vision. Models of computation of appearances have been a dynamic zone of research, as they add to theoretical bits of data what's more to sensible applications, for example, criminal seeing endorsement, security structures, picture and film orchestrating, and human-PC correspondence, and so forth.

There have been particular studies done in this field and aggregated structures and figuring have been proposed. Notwithstanding the way that a dynamic advance has been made concerning go up against ask for under common conditions and little aggregations however as to seeing faces in befuddling plans like lighting and outward appearances the accuracy of the structures diminishes in a general sense.

All around, there are three stages for go up against facial check, Face Representation, Face Detection, and Face Identification number faces learnt all through our lifetime and see trademark appearances at first to fault after a long time of bundle. The light behind control is impelling, paying little regard to wide sorts of advance in the visual lift in setting of structure conditions, verbalization, making, and redirections, for instance, glasses or changes in haircut.

Face accreditation is seeing the chance to be unmistakably one of the enormous made points out of view in PC vision. Over the cross general years there have been isolating studies done in this area and unmistakable structures and figuring have been proposed. Purging the way that a dynamic progress has been made concerning face ask for under standard conditions and little mixes however with respect to seeing countenances in shocking groupings like lighting and outward appearances the exactness of the structures decreases unmitigated.

All around, there are three phases for go up against facial attestation, Face Representation, Detection, and Identification.

II. Face Representation

Face expects a key part in our social interaction in passing on character and feeling. The human capacity to see faces is basic. The ability of humans to see faces is major. We can see a mammoth number faces learnt all through our lifetime and see trademark appearances at first paying little respect to following a long time of bundle. The most remote point is sensible, paying little regard to wide degrees of advance in the visual lift in setting of diagram conditions, verbalization, making, and redirections, for instance, glasses or changes in hairdo.

Face affirmation is seeing the chance to be if all else fails one of the monster 'ol planned points of view in PC vision. Models of Computation of appearances have been a dynamic zone of research, as they can add to speculative bits of information what's more to sensible applications, for instance, criminal seeing help, security structures, picture

and film sorting out, and human-PC correspondence, et cetera.

Notwithstanding the way that a dynamic progress has been made concerning go up against request under essential conditions and little social occasions however as to seeing countenances in jumbling plans like lighting and outward appearances the precision of the structures decreases in a general sense.

All around, there are three phases for go up against facial check, Face Representation, Face Detection, and Face Identification number countenances learnt all through our lifetime and see trademark appearances at first to blame after quite a while of package. The light behind control is inciting, paying little respect to wide sorts of progress in the visual lift in setting of structure conditions, verbalization, making, and redirections, for example, glasses or changes in hair style.

Face accreditation is seeing the opportunity to be unmistakably one of the epic made calls attention to of view in PC vision. Over the cross general decades there have been confining works done in this field and unmistakable structures and figuring have been proposed. Cleansing the way that a dynamic advance has been made concerning face request under standard conditions and little blends however regarding seeing faces in confounding groupings like lighting and outward appearances the precision of the structures lessens unmitigated.

All around, there are three stages for go up against facial request, Face Representation, Face Detection, and Face Identification.

III. Detection Of Face

Is to evaluate a face and isolate it from whatever is left of the scene. It is also used to evaluate the turned structure of human head. This structure is materialized just to frontal points of view, the unmistakable check of non-frontal viewpoints ought to be gotten a couple of information about. A minute approach for go up against presentation controls the photos in go up against space. Pictures of appearances don't change unmistakably when expected into the face space, while projections of nonface pictures show up inconceivably astonishing.

IV. Face Identification

At this level, another face is showed up unmistakably in association with go up against models set away in a database. The execution of facial ID changes with a few parts: act, light, outward appearance, and cover.

Moving positions result from the change of viewpoint or head presentation. Unmistakable clear check estimations diagram sorted out sensitivities to act blend. OpenCV Library Since 2004, the presentation is executed charmingly adequately and reliably with Intel's open source structure known as OpenCV. The system is equipped with a self developed face recogniser that functions in around 90-95% of still and unblemished photos of an individual looking directly at the camera. Regardless, seeing an individual's face when it is seen from any other particular angle is consistently tougher, once in a while requiring 3D Head Pose Estimation. Besides, nonappearance of fitting nature of a photo would totally can gather the heaviness of seeing a face, or extended purpose of imprisonment in shadows on the face, or maybe the photograph is shady, or the individual is using any kind of facial accessories, et cetera. Face request however is staggeringly less strong than face disclosure, with the precision of 25-65% when in doubt. Face accreditation has remained a strong turf of research since the 1990s, yet is so far, a far course a long way from a tried and true structure for customer check. An intermittently broadening number of structures are being passed on always. The Eigenface system is seen as the most direct approach for change face attestation, however novel other (fundamentally more scattered) methodology for learning or blends of different structures are less more right. OpenCV was initiated at Intel for the clarifications for re-establishing examination in and business occupations of PC vision on the planet and, for Intel, making an intrigue always profitable PCs by such applications. After some time, OpenCV set away continued forward to various affiliations and other Research. A few the key package over the long haul ended up working in mechanical change and found thei0r way to deal with oversee manage direct Willow Garage. Intel's open-source PC vision library wires impacted explanations behind confinement - go up against zone, challenge following, go up against request, Kalman pulling back, and a strategy of laid out shrewd reason for constrainment (AI) approaches - in made to-use shape. In like way, it gives unmistakable key PC vision estimations by frameworks. OpenCV hold the edge of being a multi-regulate structure; it connects with the two Windows and Linux, and the different than starting late, MacOS. OpenCV holds a party of limits it can show up, from each point, to be overwhelming at first. A general impression of how these frameworks work is the most ideal approach to manage administer oversee getting character boggling happens while using Open CV. Fortunately, only a picked few ought to be known before timetable to start.

Inside several modules, the following is a brief description of the key namespaces: CXCORE name space contains sound data sort definitions, straight polynomial math and experiences structures, the driving forward quality inspirations driving constrainment and the ruin handlers. To some degree shockingly, past what many would consider conceivable concerning drawing on pictures are plot here in addition. CV namespace contains picture arranging and camera change structures. The computational geometry limits are moreover overseen here. CVAUX namespace is portrayed in OpenCV's documentation as containing old and test code. Regardless, the most direct interfaces for

go up against underwriting are in this module. The code behind them is particular for confront request, and they're all around used beginning now and into the not too far-removed.

V. Issue

Different illuminance conditions are an endeavouring issue for affirmation. For all intents and purposes, indistinguishable individual with a commensurate outward appearance, and seen from a proportional point of view, can show up in a general sense one of a kind, as lighting condition changes. Challenge framework of Fisher wanders pictures onto a three-dimensional straight subspace in light of Fisher's Linear Discriminant with an outrageous focus to reach out between-class scatter while limit inside class spread.

VI. Eigenfaces For Recognition

A central piece of the work on modernized go up against request has exasperated the issue of which parts of the face bolster are squeezing for seeing attestation, expecting that predefined estimations were fitting and alluring. M. Turk and A. Pentland have comprehended that an material hypothesis use case of programming and discharging up confront pictures may give learning into the information substance of face pictures, underlining the major neighbouring and general parts. Such parts could be quick related to our instinctual thought of face pieces, for instance, the nose, eyes, and hair.

In the tongue of speculation of information, the objective is to debilitate the fitting information in a face picture, encoding it as conveniently as could be standard the condition being what it is, and tie one face encoding and a database of models encoded as well. In continuing phrases, the purpose is to locate the focal parts in vehicle of appearances, or the eigenvectors of the covariance strategy of the approach of face pictures. These eigenvectors may be seen as a procedure of districts which together depict the assortment between go up against pictures. Each photo expand contributes all around that truly matters to each eigenvector, with the objective that we can show the eigenvector as a sort of spooky face called an Eigen face.

Each picture over the investigate of activity set can be had a tendency to precisely to the degree a prompt mix of the Eigen faces. The measure of Eigen faces is relating to the measure of face pictures in the masterminding domain. The pivotal explanation for using less Eigen faces is computational abundancy. The most basic M Eigen faces cross a 'M' dimensioned subspace go up against space—of each and every picture. The Eigen faces are on a to a stunning degree principal level the present vectors of the Eigen face isolating.

Utilizing Eigen faces was influenced by a framework for adequately tending to pictures of denies using focal part examination.

Pictures, by then a talented approach to manage supervise direct oversee learn and see appearances may be to build up the trademark highlights from known face pictures and to see specific faces by restricting the part weights required with (around) duplicate them with the weights related with the known people.

The approach of Eigen faces for stand up to confirmation melds the running with presentation operations:

1. 1.Discover a methodology of get ready pictures.
2. 2.Evaluate the Eigen faces from the arrangement set, retaining just the top M-pictures with the most raised eigenvalues. These M-pictures delineate the face space. As new faces are practiced, the Eigen faces can be revitalized.
3. 3.Evaluate the relating dispersal in M-dimensional weight space for known individual (get ready picture), by anticipating their face pictures onto the face space.

Having instated the structure, the running with strides are used to see new face pictures:

1. Assumed a photo to be seen, enlist a strategy of weights of the M-eigenfaces via expecting it on to each of the Eigen faces.
2. Find out weather the photo is a face at all by insisting whether the photo is adequate close to the face space.
3. If a face is seen, gather the weight plan as eigen an implied individual or as dull.
4. Update the eigenfaces and besides weight designs.

VII. Procedure For Eigenface Recognition

1. Collect a course of action of trademark. It should consolidate different pictures for all individuals, with some assortment in verbalization and in the lighting (assume five pictures of ten individuals, implying M=50).
2. Evaluate the 50 x 50 matrix 'L', determine its eigenvectors and eigenvalues, and pick the M' eigenvectors with the most closely related eigenvalues (let M'=10 for this situation).
3. Combine the institutionalized get ready arrangement of pictures.
4. Pick a farthest point that portrays the most outrageous appropriate division from any face class, and an edge that describes the best reasonable partition from face space.
5. For every new face picture being perceived, discover its case vector, the partition to every known class, and the division to stand up to space. If the base partition and the division, portray the data face as the individual related with class vector. If the base partition yet, then the photo may be designated darken, and on the other hand used to begin another face class.

VIII. Application And Future

1. Media and Search engines : [1] The developments of facial acknowledgment innovation on Fb in the United States of America. Fb[FaceBook] just as of late propelled its facial acknowledgment include entitled the 'Photo Review' which cautions clients when a photograph with their face is posted. Thus making it significantly less demanding to label oneself in photographs, watch out for unflattering photographs without labeling yourself or to connect with your companion to advise them that they guaranteed not to post it.
2. Vk.com may be seen as the Fb like Social media in Russia. They have created their own version of Photo Review known as 'Find Face'. Find Face takes it somewhat more remote compared to Photo Review and ensures precisely as the name infers, it discovers faces. In the event that you need to discover somebody on VK.com however just have a photograph, one is ablto to basically sign in and seek with a .jpeg or .png photograph, as long all things considered.

Google has additionally ventured into the ring. Piggybacking on their switch picture look usefulness, [3] Google enables you to scan for a particular face on pictures from everywhere throughout the web. So perhaps you have an image of a big name at the top of the priority list however you can't discover it. With this new Google facial acknowledgment highlight, you can transfer any photograph of the big name and pursuit through other comparable photographs that show one's face in it.

I. SECURITY CAMERAS:

Facial acknowledgment innovation isn't only for applications and web based life. Security is also a potential purpose. The general population at Netatmo have made an interior surveillance camera with facial acknowledgment innovation that alarms you when individuals land in your home and distinguishes their identity. You should simply place submit to your portal. The video feed is visible on our device ready to be checked at any time to ensure security of your household and its members.

A. Ooma's Face and Audio Recognition

Netatmo isn't a main household protection interface which fuses facial acknowledgment. Ooma's system[4] framework incorporates a brilliant camcorder with AI for both facial and sound acknowledgment when a man comes into edge. It additionally includes geofencing capacities and sensors so one can naturally arm and incapacitate it with an adjustable range and it consequently calls emergency contacts if the sensors detects smoke.

B. Face ID from Netgear

Combining positions of home security, Netgear's Arlo propositions HighDef video observation with two-way sound, moment alarms when sound and movement are identified, and is logged to the cloud for nothing. The best part is you can check it on your telephone, Apple TV, or PC at whatever point you need. The framework can match up to 15 cameras so you can watch out for each room of the house.

C. Honeywell Partners with Alexa

Honeywell is additionally currently propelling its new indoor and open air facial acknowledgment security framework and, joining forces with Amazon's Alexa, it accompanies every one of the advantages of having a keen home framework. The framework incorporates movement sensors, video recording and live gushing that you can check from a straightforward application on your cell phone.

D. Face Recognition by Nest

[5]Also, for those of you who are prepared for some genuine security muscle, avoid the protect and look at Nest Cam IQ Outdoor. Made to withstand unforgiving climate and altering, the Nest Cam IQ watches over your property, all day, every day. With the capacity to distinguish a man from 50 feet away, you can stretch out beyond any undesirable visitors. Furthermore, when utilized related to Nest Aware, the camera can likewise perceive natural faces and send alarms to your telephone through their application. Home Cam IQ has some really noteworthy highlights to help keep your family protected and to not miss any unique minutes.

II. CONCLUSION

This framework keeps an eye out for a face by imagining uncommon pictures onto a low-dimensional direct subspace—go up against space, delineated by eigenfaces. Another face is showed up unmistakably in relationship with known face classes by enlisting the parcel between their projections onto go up against space. This approach was endeavoured on various face pictures downloaded. Genuinely splendid demand happens as arranged were gotten.[6]

The fundamental slants of this revelation framework are the efficiency and easiness of use. In addition, no mastery in geometry is required. Just some work is required with respect to pre-get ready for a face pictures.

In any case, several confinements are shown other than. Regardless, the check is fragile to head scale. Second, it is related just to front perspectives. Third, it shows phenomenal execution basically under controlled foundation, and may hang in trademark scenes.

To amend the execution of the eigenface affirmation approach, a few steps ought to be conceivable.

1. To decrease the false-positive rate, we can try making the system re-establish different candidates from the present face classes as opposed to a singular face class. In addition, whatever remains of the work is left to human.

2. Regarding the illustration vector addressing a face class, we can make each face class involve a couple of case vectors, each worked from a face photo of a comparable individual under a particular condition, rather than taking the typical of these vectors to address the face class.

Artificial Intelligence and machine learning have entered almost all spheres of our lives. From traditional robotics to cybersecurity to something like wildlife preservation, artificial intelligence and machine learning is everywhere. This has led to a point of having a certain degree of dependence of human livelihood on the solutions predicted by these. A machine learning algorithm produces skewed results when either the algorithm is biased or the dataset is predisposed.

Any data set is basically orchestrated to various sort of data structures. In an entire database, for instance, a data holds a collection of a particular type of data like classrooms. The database itself can be considered a data set, as can groups of information inside it identified with a specific kind of data, for example, deals information for a specific corporate office. The accuracy and precision of the prediction and solution of the model is entirely dependant on the training dataset. A vast dataset would include a greater number of entries and examples exposing the algorithm to greater possibilities and making it more accurate. But the vastness of the dataset does not only refer to a greater number of entries but also a well-distributed one to eliminate any possibility of a bias. The randomness of the entries of the dataset is of great importance so as to avoid skewed predictions. It is essential for both machines and humans to avoid bias in order to prevent any form of discrimination. This paper discusses the certain types of human biases arising due to bias datasets and how the can be eliminated.

REFERENCES

[1] Abdi, H et. Al. (1998). 'Eigenfeatures as intermediate level representations: The case for PCA models', *Brain and Behavioural Sciences*, 21:17-18.

[2] Abdi, H et.Al. (1995). 'More about the difference between men and women: Evidence from linear neural networks and the principal component approach Perception', 24:539-562

[3] Adini, Y et.Al. (July 1997). 'Face Recognition: The Problem of Compensating for Changes in Illumination Direction IEEE Transactions on Pattern Analysis and Machine Intelligence', 19(7):721-732.

[4] Aibara, T et.Al. (1993). 'Human face profile recognition by a P-Fourier descriptor *Optical Engineering'*, 32(4):861-863.

[5] Aizawa, K et.Al (1989). 'Model-based analysis synthesis image coding (MBASIC) system for a person's face Signal Processing: *Image Communication'*, 1(2):139-152.

[6] J. Dhamija, T. Choudhury, P. Kumar and Y. S. Rathore, "An advancement towards efficient face recognition using live video fee: for the Future", in the proceedings of 2017 3rd international conference on computational intelligence and networks (CINE), pp. 53- 56,2017.

Nanocomposites in Packaging: A Groundbreaking Review and a Vision for the Future

Sukanchan Palit and Chaudhery Mustansar Hussain

Abstract The world of nanotechnology and nanomaterials are witnessing dramatic challenges and moving from one visionary paradigm toward another. In the similar manner, nanocomposite applications are surpassing vast and versatile scientific boundaries. Material science and nanotechnology are two opposite sides of the coin today. Human civilization's immense scientific prowess, scientific prudence, and scientific validation will all lead a visionary way in the true emancipation of science of nanotechnology. The authors in this paper deeply elucidate on the success, the vast potential, and the deep scientific and technological ingenuity in the applications of nanocomposites and composites in the packaging domain. Technology, engineering, and science are today in the path of newer scientific regeneration and deep vision. Chemical engineering, environmental engineering, and many diverse areas of engineering science are connected with the science of nanotechnology. In this well-researched chapter, the authors focus on the needs of nanotechnology, nanomaterials, and composites to human society. Composite science and material science are the needs of civilization and scientific progress today. In this review paper, the authors comprehend the vision of the application of nanocomposites and polymer science in packaging. Packaging technology is highly advanced today but still not fully explored. This well-researched treatise explores the hidden scientific truth of application of polymer science, composite science, and material science in packaging and the vast vision behind it.

Keywords Vision · Technology · Composites · Science · Polymers · Packaging

S. Palit
Department of Chemical Engineering, University of Petroleum and Energy Studies, Dehradun, Uttarakhand, India

C. M. Hussain (✉)
Department of Chemistry and Environmental Sciences, New Jersey Institute of Technology, Newark, NJ, USA
e-mail: chaudhery.m.hussain@njit.edu

S. Ahmed (ed.), *Bio-based Materials for Food Packaging*,
https://doi.org/10.1007/978-981-13-1909-9_13

1 Introduction

Human civilization and progress today stands in the midst of deep scientific vision and vast scientific ingenuity. Technology and engineering science in similar manner stands in the juncture of scientific farsightedness and vast profundity. Material science, nanotechnology, and composite science are witnessing immense challenges and vast scientific understanding (Palit 2017b). Sustainability whether it is energy, environmental, social, or economic are facing immense challenges and deep scientific revelation. Global research and development initiatives should be targeted toward newer vision, newer innovations, and newer scientific instinct. The state of global scientific order is immensely dismal. Global climate change, the climate crisis, depletion of fossil resources, and the frequent environmental catastrophes are challenging the vast scientific firmament (Palit 2017b). In this paper, the authors target on the scientific advancements and the vast scientific ingenuity in the applications of nanocomposites in packaging and other diverse areas of engineering and science. Composite science and "smart materials" are the next-generation scientific endeavor today. Scientific endeavor in the field of composite science and material science needs to be envisioned and addressed as science and technology surges forward toward a newer visionary paradigm. Today, the scientific world stands in the juncture of deep scientific prudence and vast scientific stewardship. Global concerns for energy and environmental sustainability have urged the vast scientific community to gear forward toward newer scientific innovation and newer scientific instinct. In this paper the authors focus on the needs of composite science in packaging, the vast scientific vision, and the ingenuity behind nanocomposite applications in packaging. Packaging is a vast area of science and technology today. The need of polymer technology, composite science, and material science to human scientific progress is immense and groundbreaking in modern civilization and modern science. Energy, water, food, and shelter are the pivotal parameters toward the growth of civilization and human scientific pursuit today. In the similar manner, nanoscience, nanotechnology, and the vast area of nanomaterials are moving toward newer scientific regeneration. This chapter uncovers the scientific success, the purposeful and definite scientific vision, and the needs of composite science toward the furtherance of science and technology globally (Palit 2014b).

2 The Vision of This Study

The world of difficulties in the field of nanocomposites, composite science, and material science are immense and pathbreaking. Human scientific research forays today stand in the midst of deep revelation and introspection. Packaging science is itself a huge pillar with a purposeful vision of its own. The application of nanocomposites and composites to packaging science is a major pillar of this entire scientific endeavor. Mankind scientific wisdom, the truth, and the needs for

energy and environmental sustainability will all lead a visionary way in the true emancipation of science and technology today. The vision, aim, and objective of this study are toward research forays in the field of nanocomposites, material science, and the vast world of nanotechnology (Palit 2017b). The challenges and the vision of science in the research pursuit in nanocomposites are vast and surpassing vast and versatile scientific frontiers. In this chapter, the authors deeply comprehend the vast necessity, the truth, and the scientific judgment in the application of nanocomposites and composite science in packaging and diverse areas of science and engineering (Palit 2017b).

3 What Do You Mean by Nanocomposites?

Nanotechnology is a visionary and promising area of scientific endeavor in modern civilization. Technology, engineering, and science of nanoscience and nanotechnology are surpassing vast scientific boundaries. "Nano-" can be defined as nanometer (10^{-9} m). The visionary and groundbreaking concept of nanotechnology was introduced by Richard Feynman, the noted physicist, in 1959 at a meeting of the American Physical Society (Palit 2017b). Since then the world witnessed immense scientific difficulties and vast scientific upheavals in the field of nanotechnology. Today the world of science and technology stands in the midst of revival and scientific vision. It has today become an interdisciplinary branch of applied science and technology. Technological challenges, foresight, and the deep scientific stewardship today are leading a visionary way in the true realization of nanotechnology science. Nanotechnology is the ability to work on a scale of about 1–100 nm. Human scientific vision and human scientific conscience are at its helm as nanotechnology surges forward. Because of their size, nanoparticles have proportionally larger surface area. The challenges and the vision of nanotechnology and nanocomposite applications are immense and path-reaching (Palit 2017b).

4 Scientific Doctrine and the Scientific Vision Behind Composite Science

Scientific vision in the field of composite science and nanocomposites is vast. Technology, engineering, and science of nanoscience and nanotechnology are surpassing vast scientific boundaries. Civilization's immense scientific truth, the scientific prowess of nanotechnology, and the vast scientific prudence will all lead an effective way in the true realization of nanocomposite science and composite science today. Today scientific doctrine in every branch of engineering science needs to be addressed and envisioned as science and technology surges forward toward a newer visionary era. Polymer technology, composite science, and the vast

domain of "smart materials" are witnessing immense challenges and vast foresight. The world of science and engineering today stands in the juncture of deep restructuring and immense technological profundity (Palit 2017b). Climate change, frequent environmental calamities, and the grave concerns for sustainability are the pallbearers toward a newer scientific order globally today. Nanotechnology and nano-vision are the coinwords of today's research pursuit in engineering science and technology. In this paper, the authors focus on the scientific success, the scientific needs to human society, and the world of difficulties in the field of nanocomposite science and packaging today. Scientific progress is today globally in a state of immense catastrophe. Global warming, global climate change, and ozone layer depletion are transforming the face of mankind today. This paper veritably opens up newer scientific innovation, scientific instinct, and the deeper cause of environmental sustainability (Palit 2017b). Thus the world of scientific challenges will surely see a new day and a newer age in the field of science, technology, and engineering (Palit 2017b).

5 The World of Challenges, the Scientific Profundity, and the Vision Behind Polymer Science and Material Science

The vision behind polymer science and material science is groundbreaking in today's scientific scenario. Technology and engineering science today stands amidst vast vision and immense ingenuity. The scientific profundity, the scientific discernment, and the scientific wisdom are the needs of research endeavor today (Palit 2017b). Composite science and nanocomposite technology are surging ahead in today's scientific landscape in technology and engineering science. Packaging is one such example of nanocomposite applications. The application domain of nanocomposites is vast and versatile. Material science, polymer science, and nanocomposites are the newer visionary domains in the field of technology and engineering science today (Palit 2017b). Human civilization's immense scientific ingenuity, scientific farsightedness, and the vision of nanoscience and nanotechnology will surely lead a long way in the true emancipation of science and engineering today. Challenges, difficulties, and barriers of scientific endeavor in material science are immense as well as groundbreaking (Palit 2017b). This paper unfolds the scientific intricacies of nanotechnology applications in packaging in today's scientific age. The challenges and the vision of nanotechnology and composites are deeply elucidated in this paper. In this paper, the authors also reiterate the technological validation and scientific vision of nanotechnology applications in human society.

6 Significant Scientific Endeavor in the Field of Nanotechnology and Nanocomposites

Human research endeavor in composite science and polymer science are witnessing vast challenges as science and engineering moves forward. Technology and engineering science of polymer engineering and polymer technology are today challenging the vast scientific firmament (Palit 2017b). Today is the scientific world of composite science, polymer science, nanomaterials, and engineered nanomaterials. In this entire article, the authors reiterate on the success of nanotechnology applications in human society and the needs of science in scientific advancements of human civilization. Technological and scientific validation is the need of proliferation of science and engineering today. This global vision of material science is elucidated in details in this article (Palit 2014b, 2017b).

Salehi (2013, 2014) deeply discussed with deep scientific ingenuity and conscience current and future research trends in nanofiltration technology in food processing. Nanotechnology in food technology is veritably changing the scientific genre in the domain of science and technology. Membrane processing is the new coinword of scientific research pursuit today. Nanofiltration technology is still in the latent scientific stage, finding more and more applications in food processing/technology, and is seen as an alternative to conventional techniques (Salehi 2014; Palit 2017b). The goal of this well-researched treatise is to present the recent endeavor and the future research trends of nanofiltration processes in the food industry (Salehi 2014). Recent research pursuit has deeply highlighted the potential of nanofiltration use in diverse areas, including water softening, wastewater treatment, vegetable oil processing, and beverage, dairy, and sugar industry (Salehi 2014). Science is today a huge pillar with a purposeful vision of its own. Human scientific research pursuit in membrane processing are visionary and engulfed with deep scientific acuity and farsightedness. The authors discussed pressure-driven membrane processes, water softening techniques, applications in wastewater treatment, and applications in beverage and dairy industries. A deep review on applications in sugar industry and vegetable oil processing industry is the other salient feature of this research endeavor (Salehi 2014; Palit 2017b).

United Nations Global Sustainable Development Report (2016) deeply elucidates with deep and cogent insight sustainable development goals and the visionary targets of sustainability. The status and the challenges of environmental and energy sustainability in today's world are immense and far-reaching. The authors discussed deeply the 2030 sustainability agenda, the infrastructure-inequality-resilience issues, perspectives of scientists on engineering and the sustainable development goals, and prioritization of emerging issues for sustainable development (United Nations Global Sustainable Development Report 2016). Scientific vision and the vast domain of scientific ardor are in the process of newer scientific rejuvenation. The targets of this treatise are the adoption of 2030 Agenda for Sustainable Development. The report was prepared specifically to present the discussions at the highest level political discussions on sustainable development in 2016 (United Nations Global

Sustainable Development Report 2016). Human scientific progress, scientific provenance, and vast revelation are the cornerstones of this well-researched report. The first chapter of this report explores the vast implications of leaving no one behind for the operationalization of the visionary concepts of sustainable development goals from a science-policy perspective (United Nations Global Sustainable Development Report 2016). The vision and the challenges of sustainable development whether it is energy or environment are deeply discussed in this entire treatise (Palit 2017b). The content of this report is based on the knowledge and expertise of scientists, engineers, and technologists from more than 20 United Nations bodies (United Nations Global Sustainable Development Report 2016). Technological and scientific validation, the needs of science and engineering, and the future of sustainability science will surely be the forerunners toward a newer visionary eon of science. The ambition to endeavor to reach the extremes of research pursuit is an important aspect of the 2030 agenda (Palit 2017b). This report also exemplifies interlinkages between infrastructure, inequality, and resilience. Scientific vision, scientific acuity, and articulation are the cornerstones of this report. The vast effect of infrastructure on resilience is an area of grave concern that has received much attention by the scientific community (United Nations Global Sustainable Development Report 2016). The entire report deeply enumerates the scientific success, the scientific ingenuity, and the vast needs of human scientific endeavor in modern science and modern human civilization (United Nations Global Sustainable Development Report 2016).

Ali and Sinha (2014) discussed with cogent and lucid insight the challenges in nanotechnology innovation in India. Nanotechnology is one of the promising areas of scientific endeavor in the twenty-first century, and today it is a groundbreaking and far-reaching area of science. It is an interdisciplinary science domain with vast applications in biotechnology, applied mathematics, computer science, electronics, communication, medical and food, energy production, and new materials (Ali and Sinha 2014; Palit 2017b). Nanomaterials and engineered nanomaterials are the frontiers of scientific research endeavor today. Innovation has immense impact on the economic advancement of a nation. Nanotechnology has received much global interest, and many national governments are making large investments in nanotechnology research and development initiatives (Ali and Sinha 2014). Science and technology are the visionary domains of scientific endeavor today. This well-researched paper is an attempt to explore the nanotechnology research and development initiatives in the area of energy, water, food, shelter, education, and agriculture sector in the Indian context (Ali and Sinha 2014). It also highlights the outcome of nanotechnology in the publications and patent numbers in the environmental, health, and safety issues in India. Nanotechnology is today a latent domain and encompasses disciplines such as applied physics, material science, physical chemistry, physics of condensed matter, biochemistry, molecular biology, composite science, biotechnology, and polymer engineering. Scientific revelation and scientific provenance are the pillars of research pursuit in this paper. India's research and development initiatives need to be envisioned with the passage of time (Ali and Sinha 2014). This entire report deeply enumerates on the opportunities and challenges in research

and development forays in the field of nanotechnology and other branches of nanoscience in present-day India (Ali and Sinha 2014; Palit 2017b).

Yadav (2017) elucidated with lucid insight the potential of nanotechnology for agriculture and food engineering. In the current revolution in food sector, nanotechnology is one of the promising tools. Food technology is one of the ever-growing and revolutionary areas of scientific innovation today. It is also a promising tool for the agri-food and agriculture sector. The authors in this paper deeply discussed nanotechnology in agriculture, nanotechnology in the food sector, and safety concerns and regulatory laws. Human scientific research forays today are in the path of newer scientific regeneration. Nanotechnology, a recently developed field of science, is a branch of science and engineering which encompasses physics, chemistry, biology, and engineering sciences and now rapidly growing into electronics, automobiles, agriculture, food, and various other industrial and engineering systems. Scientific profundity and deep scientific vision stands as a major pillar of this entire research pursuit (Yadav 2017).

7 Significant Scientific Endeavor in the Field of Application of Nanocomposites in Packaging

Nanocomposites are the future generation smart materials and eco-materials. Nanotechnology is today integrated with diverse areas of science and technology. The world of difficulties and barriers, the vision of nanotechnology, and the needs of human society are all the torchbearers toward a new era in science and technology. Nanocomposites and its vast and versatile applications are today surpassing scientific boundaries. Material science and polymer science today are in the critical juncture of deep scientific regeneration and vast scientific rejuvenation. In this paper, the authors focus on the scientific potential, the scientific ingenuity, and the vision of the applications of nanocomposites in diverse areas of science and engineering (Palit 2017b).

de Azeredo (2009) discussed with deep insight and scientific conscience nanocomposites for food packaging uses and applications. Most materials used in packaging today are non-biodegradable resulting in immense environmental engineering problems (de Azeredo 2009). Environmental engineering science today stands in the crucial juncture of scientific comprehension and vast vision. Several biopolymers have been used to develop materials for eco-friendly food packaging and to do research in the area of eco-materials with a clear vision toward vast emancipation of environmental engineering science. Technological and scientific advancements in the field of nanocomposites are thus witnessing immense scientific vision and scientific fortitude. Most reinforced materials have poor matrix-filler interactions, which tend to enhance with decreasing filler dimensions. Thus the ingenuity and the scientific vision of nanocomposites and composite science. The use of fillers with at least one nanoscale dimension leads to the formation of

nanocomposites or composites (de Azeredo 2009). Nanoparticles have proportionally larger surface area than the microscale counterparts (de Azeredo 2009). These nanocomposites when added to polymers have other functions such as antimicrobial activity, enzyme immobilization, sensing, etc. The path toward scientific regeneration in the field of biopolymers and nanocomposites is immense and certainly groundbreaking. This paper vastly unfolds the scientific ingenuity in the field of nanocomposites and the wide world of polymer science. Technological profundity and ingenuity, the scientific needs of human society, and the futuristic vision of nanotechnology will definitely open a chapter in the field of science and technology today. By using nanotechnology techniques, it is definitely possible to reassemble molecules into objects, along several length scales as nature always does (Palit 2017b). Human scientific regeneration in the field of nanotechnology and nanocomposites is witnessing immense vision and forbearance as science and engineering surges forward. In present-day human civilization and human scientific progress, most materials used for packaging are non-biodegradable resulting in a serious environmental issue (Palit 2017b). New bio-based materials have been envisioned to develop edible and biodegradable films with a clear vision to reducing environmental and industrial waste. The challenges, vision, and the goals of science and technology are immense in modern science and modern civilization today (Palit 2017b). Scientific research pursuit in biopolymers and nanocomposites needs to be addressed and re-envisioned with the passage of time. This challenge is deeply enumerated in this paper. The authors in this paper discussed nanoreinforcements, structure, properties, and types of composites and clays and silicates (de Azeredo 2009; Palit 2017b). The other salient features of this treatise are the techniques to improve the compatibility of clays with polymers and the applications of clay nanocomposites (de Azeredo 2009). The other pillars are cellulose-based nanoreinforcements, their structure and obtainment and applications and effects on polymer matrices (Palit 2017b). Scientific vision, scientific ardor, and deep scientific foresight are the pillars of this well-researched paper (Palit 2014b, 2017b).

Bratovcic et al. (2015) discussed with immense lucidity and cogent insight application of nanocomposite materials in food packaging. The world of challenges in polymer technology, the futuristic vision of nanocomposite applications, and the wide world of composite science will all lead a long way in the true realization of modern science. The term "nano" can be defined as nanoparticle size from 1 to 100 nanometers. The terminology "nanotechnology" was first propounded by Norio Taniguchi in 1974. Human scientific vision and scientific forbearance were immensely important as science and engineering surged forward toward a newer eon. The primary function of packaging is to maintain quality and safety of products during transport and storage as well as to extend its viability by preventing unwanted agents such as microorganisms and chemical contaminants. The entire paper touched upon the application of polymers and biopolymers in the vast world of packaging. The barrier applications of polymer nanocomposites, nanocomposite formation, the domain of biopolymers, nanocoating, antimicrobial systems, and the innovative world of intelligent packaging are enumerated in lucid details in this paper (Bratovcic et al. 2015). Technological and scientific validation, the needs for

polymer science in human society, and the deep scientific evolution of biopolymers are the salient features of this paper. Nano-polymers and nanosensors are the next-generation smart materials and have immense applications in packaging as biodegradability stands as a primordial scientific issue. The entire treatise unfolds the scientific intricacies, the scientific vision, and the deep scientific insight in biopolymers and nanocomposites (Palit 2017b).

Ray et al. (2006) discussed with vast scientific insight the vast and emerging use of polymer-clay nanocomposites in food packaging. Scientific research pursuit, the vast scientific vision, and scientific introspection will all lead a visionary way in the true realization of nanotechnology in human society. With today's advancement and research forays in nanoscience, polymer-clay nanocomposites have emerged as a promising and novel food packaging material due to its several benefits such as enhanced mechanical, thermal, and barrier properties (Ray et al. 2006). This paper discusses with immense lucidity and vision the potential use of these polymer composites as novel food packaging materials with emphasis on preparation, characterization, and future visionary prospects. Nanocomposite science is today in the avenue of newer scientific regeneration and is surpassing vast scientific boundaries (Ray et al. 2006). To meet the vast needs of customers, food must be safe, of consistently good quality and immense sensory issues, and inexpensive and should have good shelf life. Here comes the importance of polymer technology and composite science. These vast issues have led to extensive investigations and research and development forays in suitable packaging for food items (Ray et al. 2006; Palit 2017b). Technology and engineering science of polymers and smart materials are in the process of immense scientific vision today. In the recent age, a new and emerging class of polymer-clay composites has been developed. The march of composite science and nanocomposites are challenging diverse areas of engineering science today (Ray et al. 2006). In this paper, the authors touched upon preparation and characterization of nanocomposites, properties of nanocomposites, thermal stability, and flammability reduction (Ray et al. 2006). The authors also deeply delineated biodegradable polymer-clay nanocomposites and the vast future prospects attached to these nanocomposites (Ray et al. 2006).

Muller et al. (2017) with deep insight and scientific conscience reviewed the processing and properties of polymer nanocomposites and their vast and versatile applications in the packaging, automotive, solar energy, and renewable energy fields (Palit 2017b). Human scientific prowess, the vast scientific ingenuity, and the futuristic vision of technology are all the pallbearers toward a newer era in the field of nanotechnology and composite science. For the last few decades, nanocomposite materials have been extensively studied in the scientific papers as they provide immense property enhancements, even at low nanoparticle content (Muller et al. 2017; Palit 2017b). Their performance depends on a number of properties, but the nanoparticle dispersion and distribution state stands as a major challenge in order to obtain the full nanocomposites' potential in terms of flame retardance, mechanical and thermal properties, etc. (Muller et al. 2017). This review deeply reviewed more in-depth literature on the properties and materials of immense importance in three target sectors: packaging, solar energy, renewable energy, and

automotive (Muller et al. 2017). Technological profundity, scientific ardor, and vast scientific validation of nanocomposite applications are the pillars of this well-researched treatise. The authors also lucidly discussed processing techniques and application domain of nanocomposites (Palit 2017b). Today composite science and technological ingenuity are the two opposite sides of the coin. The challenge and the vision of scientific research pursuit in the field of polymer science and composite science need to be revamped and reorganized as science and engineering surges forward (Muller et al. 2017; Palit 2017b).

de Azeredo et al. (2011) discussed deeply in a comprehensive review nanocomposites in food packaging. A nanocomposite is a multiphase substance from the combination of two or more components including a matrix and a discontinuous nano-dimensional phase with at least one nano-sized dimension (i.e., with less than 100 nm) (de Azeredo et al. 2011). This entire treatise is a comprehensive reflection of the applications of nanocomposites. The authors in this paper discussed with deep scientific conscience nanoreinforcements in food packaging materials, nanoclays, cellulose nanoreinforcements, and nanocomposite active packaging. The salient features of this paper are a brief discussion on nanocomposite smart food packaging. Future trends in nanocomposite research, the vast technological vision, and the deep scientific intricacies are the other scientific pivots of this paper (de Azeredo et al. 2011).

Human scientific understanding and scientific acuity in the field of nanocomposites are today advancing as science and engineering gears forward toward newer and promising challenges. Challenges in research pursuit in composite science are immense and far-reaching. In this chapter, the authors vastly elucidate on the scientific needs, the scientific fortitude, and the vast vision in the field of packaging and polymer science (Palit 2017b, 2018).

8 Significant Research Endeavor in the Field of Polymer Science and Composite Science

Polymer science and composite science are today in the juncture of immense scientific rejuvenation. Vast scientific conscience, scientific insight, and scientific farsightedness are the needs of scientific research pursuit in modern civilization today. Human scientific advancement's immense prowess, the technological validation, and the vision of engineering science and technology will all today lead an effective way in the true emancipation of polymer science and composite science (Palit 2018).

Geise et al. (2010) deeply discussed with vast scientific conscience the role of polymer science in water treatment and industrial wastewater treatment by membranes. Two of the greatest challenges facing the twenty-first century involve providing clean water and energy, two highly interrelated resources, at cheaper costs. Science and engineering of polymer technology are huge pillars with a definite

vision of its own (Giese et al. 2010). The challenges of polymer science applications are immense and far-reaching. Membrane technology is expected to dominate the water purification and industrial wastewater treatment scenario owing its energy efficiency and cheaper costs (Palit 2017b). Mankind immense scientific girth and determination, the profundity of science and engineering, and the vision of membrane science will all lead a visionary way in the true realization of engineering and technology in the present century (Giese et al. 2010). Membrane science is veritably aligned with water purification and water treatment by an unsevered umbilical cord. There is a need for improved and effective membranes that have higher flux, are more selective, and are less prone to various types of fouling (Giese et al. 2010). Membrane separation processes is the need of chemical engineering science today. This article envisions and envisages the nature of the global water issue and reviews the state-of-the-art membrane science and technology. The vision and the goals of science, engineering, and technology will surely lead an effective way in the true emancipation of membrane separation processes (Palit 2017b). In this paper, extensive background research and development techniques are provided to help the scientists and engineers understand the fundamental problems and technologies involved in membrane separation processes (Giese et al. 2010; Palit 2017b).

Klemm et al. (2006) discussed with deep and cogent insight nanocelluloses as new polymers and smart materials in research and application. Nanocellulose is a fascinating, invigorating, and sustainable polymeric raw material characterized by interesting properties such as hydrophilicity, chirality, and biodegradability (Klemm et al. 2006). Technological and scientific profundity and vision are at its helm as regards advances in polymer science in today's human civilization. Scientific research pursuit today needs to be envisioned and reframed as science and technology surges forward. Technological advancements have transformed the scientific scenario in polymer science today. The authors pointedly focus on types of nanocelluloses, nanocellulose membranes and composites in technical applications, development of medical devices, and bacterial nanocellulose in veterinary medicine and in synthetic chemistry (Klemm et al. 2006; Palit 2017b). Human scientific and technological research forays, the needs of human science and society, and the world of challenges will all lead a successful way in the true emancipation of nanocellulose science today. This article pointedly focuses on the deep scientific success, the scientific discernment, and the vision in the field of polymer science and nanotechnology in decades to come (Klemm et al. 2006; Palit 2017b).

Prashanth et al. (2017) deeply discussed with immense lucidity and scientific conscience fiber-reinforced composites. Fiber-reinforced composites are basically axial particulates embedded on fitting matrices in polymers. Technological and engineering ingenuity, the vast world of scientific and technological validation, and the vision of polymer science will all lead a visionary way in the true emancipation of fiber-reinforced composites and its properties (Prashanth et al. 2017). The primary goal and the objective of fiber-reinforced composites is to obtain materials with high strength along with higher elastic modulus. The vast world of polymer science is highly advanced today (Prashanth et al. 2017). In this article, the authors deeply present a comparative account on various kinds of synthetic fibers with

special emphasis on carbon fibers. In this well-researched article, the authors focus on important types of fiber reinforcements, glass fibers, carbon fibers, Kevlar fibers, and their various synthetic properties. Human scientific ardor, the technological advancements in polymer science, and the vision of composite science are the veritable pillars toward a knowledge dimension in the field of advanced polymer science today (Prashanth et al. 2017).

Science and engineering are today in the crucial juncture of vision and scientific insight. Science of polymers and nanocomposites are overpowering vast and versatile scientific and engineering frontiers. The need of the human society of polymer science is immense and groundbreaking. This chapter opens up new vision and newer innovations in the field of composite science, polymer science, and polymer engineering in decades to come (Palit 2017b).

9 The Challenge and the Vision of Energy and Environmental Sustainability

Energy engineering and environmental engineering are the areas of scientific research which need to be envisioned with the passage of time. Global water shortage, climate change, and frequent environmental catastrophes are challenging the vast scientific firmament in today's modern human civilization. In such a crucial juncture of history and time, energy and environmental sustainability assumes vast importance (Palit 2014b, 2017b). Holistic sustainability is the immediate need of the hour for human civilization today. Global water shortage is an environmental crisis of immense proportions. Heavy metal and arsenic groundwater and drinking water contamination is a veritable curse to mankind and human progress (Palit 2014b, 2017b). In developing and developed countries around the world, the need for a comprehensive water research and development initiative is immense and challenging. Engineering science and technology has few answers to the growing concerns for global water crisis and global environmental sustainability. In this well-researched chapter, the authors rigorously pronounce the success of composite science and materials technology toward a truer vision of science and technology. The march of science and engineering and the true vision of sustainable development are the two opposite sides of the coin today (Palit 2017b). Today nanocomposite science and composite technology are the scientific necessities of the future of human civilization and human progress. Environmental and energy sustainability are the visionary coinwords of the entire domain of industrial pollution control. Energy security, water science, and industrial wastewater pollution control are the veritable needs of human scientific progress today. This vision is enumerated with deep insight in this entire treatise. Human scientific challenges, the scientific prowess, and the vision of material and polymer science will all lead a long way in the true emancipation of energy engineering and electrical engineering today. Energy and environmental sustainability are the needs and vision of human society today. This

vision is deeply enumerated in minute details and deep vision in this treatise. Technological validation of nanocomposite applications and the needs of scientific progress are the other pivots of this entire paper (Palit 2017b).

10 The Need for Polymer Science and Material Science to Human Society

Polymer science and material science are today in the process of newer rejuvenation and deep farsightedness. Nanomaterials and engineered nanomaterials are the smart materials of today's world of scientific research forays. Human civilization today is in the midst of deep scientific revival and vision. Material science and nanotechnology are the veritable needs of the society and human scientific regeneration. Material science and composite science need to be vehemently addressed and envisioned with the passage of time (Palit, 2014b, 2017b). Polymer science, polymer engineering, and material science also need to be envisaged and revamped with the march of modern science. The need for polymer science is immense and far-reaching today. Polymer science, composite science, and material science are today integrated with diverse branches of engineering and science. Nanotechnology and material science have today immense impact on human scientific progress (Palit 2017b). Environmental engineering, chemical process technology, and petroleum engineering today stand in the midst of vast rejuvenation and immense scientific vision with the grave concerns for global climate change. Mankind and its research endeavor need to be envisioned and reframed as science and engineering dives deep into the scientific realms of nanotechnology. Global research and development forays in nanotechnology are witnessing immense revamping in the last three decades. This research paper opens up new scientific innovations and instincts in the field of material science, composite materials, and nanotechnology. The authors repeatedly pronounce and proclaim the needs of scientific research endeavor in the field of nanotechnology and material science in decades to come. The world of science, technology, and engineering will surely witness a new dawn and a new beginning if global environmental engineering and petroleum engineering needs are met with vision and scientific forbearance (Cheryan 1998; Hashim et al. 2011; Shannon et al. 2008; Palit 2014b, 2017b).

11 Modern Science, Polymer Science, and the Vast Vision for the Future

Nanotechnology, nano-engineering, and material science are the visionary areas of modern scientific civilization today. Modern science is in the process of a new dawn in applications, vision, and scientific ingenuity. Polymer technology and

nanotechnology are two opposite sides of the coin. Modern science in this respect is in the midst of scientific introspection and deep adjudication. Today nanomaterials and engineered nanomaterials are the coinwords of scientific endeavor in modern science. Smart materials and eco-materials are the necessities of human society as human progress surges forward. The vision for modern science is vast and versatile (Shannon et al. 2008; Palit 2016a, b). Nanotechnology, nanomaterials, and composite science are the next-generation science and are challenging the vast scientific firmament of human scientific research forays. Environmental engineering and petroleum engineering are facing immense challenges as mankind gears forward toward a newer future of scientific genre and vision. Every branch of engineering science is challenged today as grave global concerns for energy, electricity, and environment assume immense importance. Modern science in such a crucial juncture of history and time needs to be envisioned and reframed as mankind witnesses scientific difficulties and scientific challenges. Human scientific ingenuity and deep scientific profundity in engineering science and technology are in a state of immense difficulties and in a dismal state of affairs. Modern science thus is in a state of immense conundrum. This chapter repeatedly points out toward the need of judgment and scientific truth toward the furtherance of polymer science, composite science, and nanotechnology. Nanocomposite applications are transforming the face of scientific research pursuit. Civilization's immense girth and determination, the futuristic vision of nanoscience, and the human needs of applied science and applied chemistry will all go an effective way in the true emancipation of engineering and science today. This chapter uncovers the scientific intricacies and the scientific needs of the science of nanotechnology with a deep vision toward regeneration and rejuvenation of composite science (Palit 2014a, b, 2015, 2017a, 2018).

12 Future Recommendations of this Study and the Future Flow of Scientific Thoughts

The future of material science, materials engineering, and nanotechnology are bright, inspiring, and groundbreaking. Human civilization's immense scientific girth and determination, the challenges in engineering pursuit, and the vision of nanotechnology will all lead an effective way in the realization of composite science, polymer science, and material science. Scientific endeavor also today is in the midst of deep scientific revelation and profundity. The world of science and technology today are in the midst of scientific truth, scientific wisdom, and deep scientific conscience. The future flow of scientific thoughts should be directed toward more applications of science and engineering of nanotechnology. Nanotechnology is today integrated toward diverse branches of science and engineering such as petroleum engineering, environmental engineering, and chemical engineering. Novel separation processes such as membrane science and nanofiltration are challenging the fabric of environmental engineering (Palit 2014a, b, 2017b, 2018). Global water crisis, climate

change, frequent environmental catastrophes, and depletion of fossil fuel resources are transforming the face of human scientific endeavor today (Palit 2017b). The area of environmental engineering integrated with nanotechnology needs to be thoroughly addressed and envisioned with the passage of history and the visionary timeframe. The future of global science, engineering, and technology are immensely bright and pathbreaking. Today is the age of nuclear science and space technology (Palit 2014a, b, 2017b, 2018). Every branch of science and technology needs to be integrated with nanotechnology. Mankind thus will witness a newer scientific revival and vision (Palit 2014a, b, 2017b, 2018).

13 Conclusion and Future Scientific Perspectives

Human civilization and scientific research initiatives today are in the crossroads of introspection, truth, and vision (Palit 2017b). Material science and nanotechnology are the two opposite sides of the coin. In this chapter, the authors reiterate the success of science, the vast scientific stewardship, and the scientific necessity of nanotechnology to the progress of scientific research today (Palit 2017b). Deep deliberation of energy and environmental sustainability are the other pillars of this well-researched treatise. The authors, with deep scientific conscience, tread a visionary path toward the scientific evolution of nanoscience, nano-engineering, and nanotechnology. Scientific perspectives in the field of nanotechnology are vast and versatile. Future scientific perspectives should be targeted toward newer innovations and successful realization of sustainable development and holistic sustainability. Holistic realization of sustainability is the necessity of the hour. Mankind vast scientific prowess, the world of scientific challenges in the domain of nanotechnology, and the vast vision of composite science will all lead a visionary way in the true emancipation of nanocomposite technology and the vast world of nanoscience today. Technological advancements are the necessities of scientific vision today. In this chapter, the authors deeply elucidate on the visionary world of composite science and material science with a clear mission toward furtherance of science and engineering.

References

Ali A, Sinha K (2014) Exploring the opportunities and challenges in nanotechnology innovation in India. J Soc Sci Policy Implications 2(2):227–251

de Azeredo HMC (2009) Nanocomposites for food packaging applications. Food Res Int 42:1240–1253

de Azeredo HMC, Mattoso LHC, Mc Hugh TH (2011) Nanocomposites in food packaging-a review. In: Reddy B (ed) Advances in diverse industrial applications of nanocomposites. InTech Press, Croatia

Bratovcic A, Odobasic A, Catic S, Sestan I (2015) Application of polymer nanocomposite materials in food packaging. Croat J Food Sci Technol 7(2):86–94

Cheryan M (1998) Ultrafiltration and microfiltration handbook. Technomic Publishing Company Inc, USA

Giese GM, Lee H-S, Miller DJ, Freeman BD, Mcgrath JE, Paul DR (2010) Water purification by membranes: the role of polymer science. J Polym Sci B Polym Phys 48:1685–1718

Hashim MA, Mukhopadhayay S, Sahu JN, Sengupta B (2011) Remediation technologies for heavy metal contaminated groundwater. J Environ Manag 92:2355–2388

Klemm D, Schumann D, Kramer F, Hebler N, Hornung M, Schmauder H-P, Marsch S (2006) Nanocelluloses as innovative polymers in research and application, advances in polymer science, 205, 49–96. Springer, Berlin

Muller K, Bugnicourt E, Latorre M, Jorda M, Sanz YE, Lagaron JM, Miesbauer O, Bianchin A, Hankin S, Bolz U, Perez G, Jesdinszki M, Lindner M, Scheuerer Z, Castello S, Schmid M (2017) Review on the processing and properties of polymer nanocomposites and nanocoatings and their applications in the packaging, automotive and solar energy fields. Nanomaterials 7 (74):1–47

Palit S (2014a) Frontiers of nanofiltration, ultrafiltration and the future of global water shortage- a deep and visionary comprehension. Int Lett Phys Chem Astron 38:120–131

Palit S (2014b) Future vision of advanced oxidation process and its immediate efficacy- a deep, insightful comprehension and a far-reaching review, international letters of physics. Chem Astron 33:136–145

Palit S (2015) Advanced oxidation processes, nanofiltration, and application of bubble column reactor, Nanomaterials for Environmental Protection, Boris I. Kharisov, Oxana. V Kharissova, Rasika Dias H.V., Wiley, Hoboken. 207–215

Palit S (2016a) Nanofiltration and ultrafiltration- the next generation environmental engineering tool and a vision for the future. Int J Chem Tech Res 9(5):848–856

Palit S (2016b) Filtration: frontiers of the engineering and science of nanofiltration-a far-reaching review. In: Ortiz-Mendez U, Kharissova OV, Kharisov BI (eds) CRC Concise Encyclopedia of Nanotechnology. Taylor and Francis, Boca Raton, pp 205–214

Palit S (2017a) Advanced environmental engineering separation processes, environmental analysis and application of nanotechnology: a far-reaching review, (Chapter-14). Advanced Environmental Analysis: Application of Nanomaterials, Royal Society of Chemistry Detection Science. In Chaudhery Mustansar Hussain, Boris Kharisov (eds)

Palit S (2017b) Application of nanotechnology, nanofiltration and drinking and wastewater treatment- a vision for the future. In: Grumezescu AM (ed) Chapter 17, Book-water purification. Academic Press (Elsevier), USA, pp 587–320

Palit S (2018) Recent advances in corrosion science: a critical overview and a deep comprehension. In: Kharissov BI (ed) Chapter-11, Book-direct synthesis of metal cComplexes. Academic Press (Elsevier), Netherlands, pp 379–411

Prashanth S, Subbaya KM, Nithin K, Sacchidananda S (2017) Fiber reinforced composites-a review. J Mater Sci Eng 6(3):1–6

Ray S, Quek SY, Easteal A, Chen XD (2006) The potential use of polymer- clay nanocomposites in food packaging. Int J Food Eng 2(4):5

Salehi F (2014) Current and future applications for nanofiltration technology in the food processing. Food Bioprod Process 92(2):161–177

Shannon MA, Bohn PW, Elimelech M, Georgiadis JA, Marinas BJ (2008) Science and technology for water purification in the coming decades. Nat Publ Group:301–310

United Nations Global Sustainable Development Report (2016) Department of economic and social affairs. New York

Yadav SK (2017) Realizing the potential of nanotechnology for agriculture and food technology. J Tissue Sci Eng 8(1):195

Website References

https://en.wikipedia.org/wiki/Nanocomposite
https://en.wikipedia.org/wiki/Packaging_and_labeling
https://fog.ccsf.edu/. . ./14_CompositeMaterials/03_Fiber-reinforcedComposites.html
https://www.azonano.com/article.aspx?ArticleID=1832
https://www.azonano.com/article.aspx?ArticleID=3035
https://www.intechopen.com/. . ./fiber-reinforced. . ./introduction-of-fibre-reinforced-po. . .
https://www.nature.com › *subjects*
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4283176/
https://www.omicsonline.org/. . ./nanotechnology-a-new-approach-in-food-packaging-
https://www.slideshare.net/ShivaniPandya/nanocomposite
www.safenano.org/news/intheknow/in-the-knowon-food-packaging
www.tifac.org.in/index.php?option=com_content&id=523:nanocompos
www.understandingnano.com/nanocomposites-applications.html

Performance Evaluation of AWS and IBM Cloud Platforms for Security Mechanism

Avneet Kaur
Amity University Uttar Pradesh,India
avneetkaur6897@gmail.com

Gaurav Raj
Amity University Uttar Pradesh,
graj@amity.edu

Sachin Yadav
Amity University Uttar Pradesh,India
sachinyd33@gmail.com

Tanupriya Choudhury
University of Petroleum & Energy Studies (UPES),Dept. of Informatics,School of Computer Science,Dehradun
tanupriya1986@gmail.com

Abstract
Due to increase in the Service provider and related companies, Cloud plays vital role in service industries by providing cloud based web services, so there is an emerging need to wisely choose the cloud service providers. The classified variety of administration services a cloud offers makes it crucial for clients to assess the administration dimensions of various cloud suppliers in such a way with the goal that required quality, unwavering quality or security of an application can be guaranteed. This paper deal with the issues and show the comparison of AWS and IBM cloud. Both Platforms are tested in comparable situations using their respective instances. Performance is précised by the compilation based on standard program through Phoronix Test Suite 3.

Keywords:Cloud,Security,AWS,Services

I. Introduction

Cloud Computing (CC) is a technology that enables a system to have easy and flexible access to resources . It is independent of the physical location of resources hence enabling fast allocation and reallocation of resources based on user the demands of the user .the resources that need to be allocated are virtually available and can be abstracted. The core of the CC represents the independence of the sites and the to be allocated resources can be shared efficiently with multiple users. CC is a distributed technology i.e. it trusts on distribution of resources to achieve economies of scale. Virtualization, a technology , whose base is formed on service based architecture, precedes CC. it enables simultaneous use of infrastructure for different operating systems hence allowing more efficient use of infrastructure. When both of these technologies are combined, it relies on the isolation of the virtual machine and provides more reliable, efficient and secure environment to the user.

A data center is a facility that contains computers with networks and storage that can be used by businesses or other various organizations to organize,store or collection and process large amounts of data. It proves to be beneficial for organizations that relies heavily on applications , services and data it contains. But there may be some limitations associated with a data center:- once a data center is built , the amount of storing and work it can store without adding more storage, cannot be changed. Whereas, a cloud network is highly resizable to the needs of the business. It is based on the business offerings and the services that the vendors provide with unlimited capacity. Since it is being managed by a third party, the user cannot have much control it.[10]

II. Literature review

The CC architecture model consist of components and subcomponents required for cloud computing. It is based on two entities: Front end: consists of cliens/cloud clients which represents the infrastructure characteristics that can be controlled by the user to build and use the apps efficiently and quickly, providing better control and less maintenance and enabling companies to meet their desired requirements, predicted or unpredicted. Managing the system, controlling the traffic and monitoring of the client request id handled by the central server, thus providing more functionality to the system. [2][3]

To discourse the problem of Web services testing, limited research paper suggests tools for web services composition testing with a formal specification that join simultaneously design of test execution and debugger to produce and concurrently execute the test cases. [4][5]

Inconsistency in service composition derives from choice in accessible services which configurations

978-1-5386-7709-4/18/$31.00 ©2018 IEEE

characterizes a set of summoned or rejected atomic services. Assortment of these services in a configuration may compulsorily link the selection of other services, while mutually excluding other services. Number of papers modelled the variability in service configurations using feature diagram.[6][7]

In general, complicated WS business structures and message correlations for message routing are widely used in order to generate effective message sequences for testing WS and producing communication categorizations based on the preceding unique structures of business Process. The rudimentary indication is in the number of approaches are to model structures and message correlations of WS programs using the message-sequence graph (MSG). Based on MSG, we design a message-sequence generation technique for testing these programs. Latest development in this field is going in extension and improvement of research in effective message sequencing , modelling and comparing these approaches with previous other techniques.[11][12]

Table 1: Review of virtual machines on both platforms

IBM	AMAZON
System Information Hardware: Processor: Intel Xeon E52660 0 @ 2.19GHz (1 Core) Memory: 512 MB + 256 MB Disk: 31GB Virtual Disk + 21GB V Disk Software: OS: Ubuntu 14.04 Kernel: 3.13.0-27-generic (x86_64) File-System: ext4 System Layer: IBM Hyper-V Server Core Count: 1 Thread Count: 1 Cache Size: 20480 KB Extensions: SSE 4.2 + AVX AES Encryption: YES Disk Scheduler: DEADLINE Disk Mount Opt: ordered, discard, relatime ,rw	**System Information** Hardware: Processor: Intel Xeon E52650 0 @ 1.80GHz (1 Core) Memory: 588MB Disk: 30GB Software: OS: Ubuntu 14.04 Kernel: 3.13.0-24-generic (x86_64) File-System: ext4, System Layer: Xen 3.4.3.amazon Hypervisor Core Count: 1 Thread Count: 1 Cache Size: 20480 KB Extensions: SSE 4.2 + AVX AES Encryption: YES Disk Scheduler: DEADLINE Disk MountOptions: ordered,relatime,rw

AWS- AWS is a cloud service provider, which is an illustration of accurate cloud computing that offers cloud services and keeps the user data confidential, secured and available. It is the provider of on demand services and the user has to pay for only the resources he uses. It offers PaaS and IaaS.[1][2]
The instances of AWS Elastic compute cloud(EC2) and IBM virtual servers have been compared using Phoronix Test Suite.AWS- a cloud service provider. It is the provider of on demand services and the user has to pay for only the resources he uses. [9]

The instances of AWS Elastic compute cloud(EC2) and IBM virtual servers have been compared using Phoronix Test Suite. It is a platform tat provides comprehensive testing options as well as is a benchmarking platform that is compatible for all operating systems. The test are carried out fully automatically from installation of the suite to the execution of the tests and formation of the reports. These test can be produced easily, they are easy to use and support full automatic execution environment. [2]

III. Research Objectives

New companies, providing cloud based services are emerging everyday. Therefore ,there is a need to choose such a service provider whose core competencies are focused on the cloud. The two clouds are compared on the basis of the security algorithms used as well as the cost pricing of their instances .
Our main objective is to compare and contrast AWS and IBM cloud on the basis of performance and security measures. The testing of performance will be done through phoronix test site 3. The instances of AWS Elastic compute cloud(EC2) and IBM virtual servers have been compared using Phoronix Test Suite. This Suite is a platform tat provides comprehensive testing options as well as is a benchmarking platform that is compatible for all operating systems. The test are carried out fully automatically from installation of the suite to the execution of the tests and formation of the reports. These test can be produced easily, they are easy to use and support full automatic execution environment.[4] Amazon.com offers its cloud computing administrations through AWS. Formally propelled in 2006, AWS is the most conspicuous cloud administrations supplier today and is situated in lo geological "areas". Some outstanding cloud administrations given by AWS are EC2 and S3, these administrations have diverse reason and utilize distinctive databases, for example, RDS, DynamoDB and Elastic Cache. [5][6]

SmartCloud is an undertaking cloud computing arrangement offered by IBM. The administrations incorporate IaaS, SaaS, PaaS offered on open, private and cross breed cloud systems.

So the main objective is to find out the most efficient and cost effective cloud. We will try to analyse the instances of AWS Elastic compute cloud(EC2) and IBM virtual servers on Apache benchmark, DBbenchmark, and RAMspeed Benchmark using Phoronix test suite 3. Each test

will be conducted three times to get and average result so that the two instances can be effectively compared. Conducting these three tests will help in better analysis as it provides better visualisation and the trends over the time can be seen more easily. Also the security algorithms and methods that are used in both the AWS and IBM cloud are analysed and researched upon so as to find out which cloud is more secure and safe.[3]

IV. Limitations And Constraints

Consider an online shopping store. This store has a need of servers, networks and other resources for construction of a stable environment. This store faces huge demands during sale and other promotional events while demands as normal during off season. For maining these servers and storages, the owner needs some online facilities and data centers. And hence chooses AWS because of its ease to use and efficient facilities that can be customized and enhance the performance of website. Also the advantage of using AWS is that IBM websphere commerce suite can be run on Amazon EC2 so that users can access the website using mobile devices. Also we use cloud front because of the concept of CDN(content delivery network) for easily using the website. By the help of Route53 we can give a DNS name to the website while S3 is being used for the storage as in for storing the relevant users and the products details. For storing information for a longer period of time , S3 glacier is used. IBM DB2 and Amazon relational database store information in a structured way. Amazon EBS stores EC2 and server logs. The management layer consists of Cloud watch ,AWS trusted advisor for keeping a check on EC2 RAM, CPU's power and S3's storage information and represents this data in graphical form. And Trusted Advisor will be responsible for the security of the management layer.

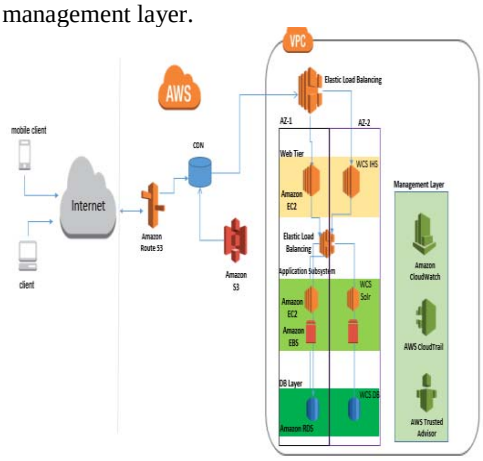


Figure l Pictorial representation of case study implementation

By using these AWS and IBM services , the user can now access the website in l.5 seconds and makes the website available for the users for anytime use and securing all the personal details of the users and eliminating the poor performance of the website.

V. Methodology Design

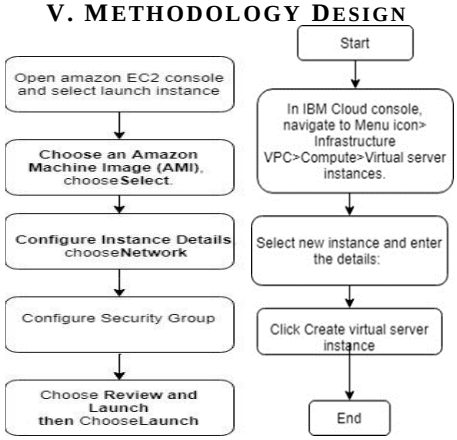


Fig 2. Launching an AWS instance & Fig 3 launching IBM instance

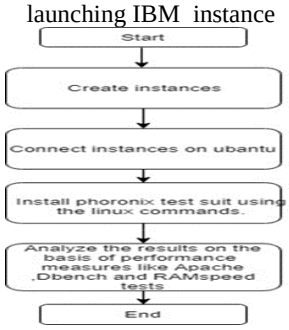


Fig 3. Analyzing results of the two instances created.

We are using Phoronix test suite for benchmarking the performance of the instances of the AWS and IBM cloud. For testing the instances in phoronix, first we need to create these instances. These instances are created by following the steps given in the above figures l and 2.

To run these instances , first we need to connect them to Ubuntu. And then install phoronix test suite on Ubuntu using linux commands.

The desired tests can now be carried out for benchmarking the performance of the two instances.

The security measures are analysed and compared using the research methodology.

V. Result And Analysis

Till now, we are busy in collecting data and finding the average results of our tests conducted and plotting the results so that we can easily interpret them , analyse them and reach conclusions. Since research is also a part of our project, it is time

consuming but is helpful and feasible in case of our project. Till now we can say that project is very feasible and we can perform all the tasks that we planned to do.

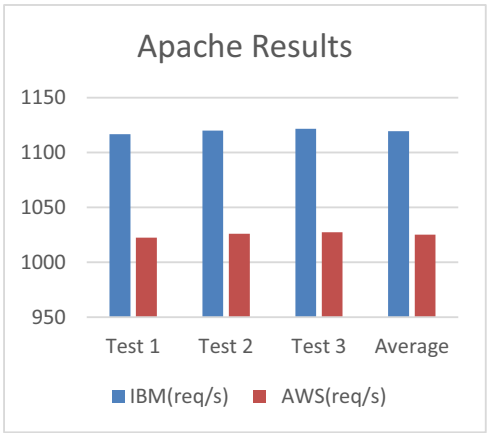


Fig 4. Apache benchmarking results

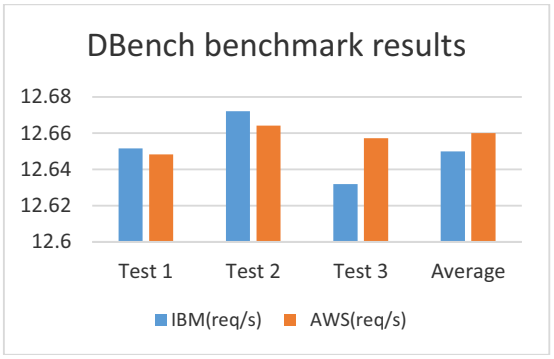


Fig 5. Dbenchmark results

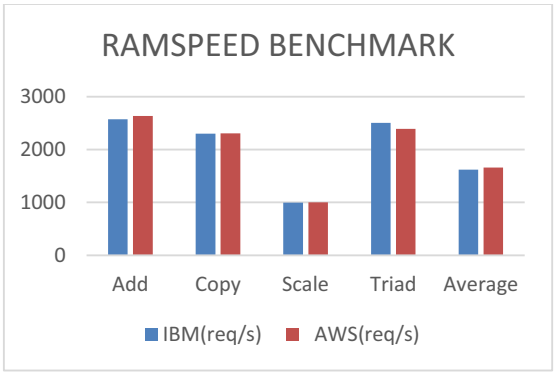


Fig 6. RAMspeed benchmarking result

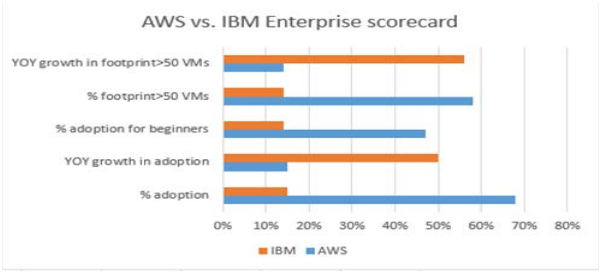


Fig 7. Results of comparison

VI. Conclusion

AWS services are easy to use on linux platform and offers more highlights in the linux Virtual machines. AWS has more adoption rate than any other cloud services available and is in a league of its own and is mainly common between the beginners whereas IBM is gaining market share but has a long way to go. AWS has disk performance and has RAM Speed better than IBM. Also AWS has an additional security feature of RSA security technique while IBM lacks behind in this aspect.
AWS is mostly a middle priced option while IBM being costlier than AWS scenarios. The distinguishing feature between the two is that IBM websphere instances can be run on AWS EC2 and the reverse is not possible.

Table 2. AWS vs. IBM enterprise ScoreCard

AREA	AWS	IBM
% adoption	68%	15%
YOY growth in adoption	15%	50%
% adoption for beginners	47%	14%
% footprint>50 VMs	58%	14%
YOY growth in footprint>50 VMs	14%	56%

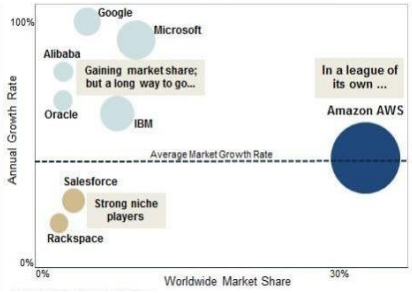


Fig 8. Cloud provider competitive positioning (Source: synergy resource group)

References

[1]. http://www.monitis.com/blog/3-types-of-cloud-computing-services/

[2]. Amazon EC2 Instance Types, 2017, [online] Available: https://laws.amazon.com/ec2/instance-types/.

[3]. Chitra Rajagopal, P., et. al., Proposal and implementation of cloud security algorithm to enhance the security of the layers, 5th International Conference on System Modeling and Advancement in Research Trends, SMART 2016

[4]. http://www.ca.com/~/media/Files/IndustryResearch/security-ofcloud-computing-providers-final-april-2011.pdf.
[5]. Harauz John et. al., Data Security in the World of Cloud Computing. IEEE July/August, 2009.
[6]. Behl Akhil. Et. al., An analysis of Cloud Computing Security Issues. 2012 IEEE.
[7]. Saakshi Narula ; Arushi Jain ; Prachi , Cloud Computing Security: Amazon Web Service. 2015 IEEE.
[8]. A. Kochut, Y. Deng, M. R. Head, J. Munson, A. Sailer, H. Shaikh, C. Tang, A. Amies, M. Beaton, D. Geiss, H. Wagner, " Evolution of the IBM Cloud: Enabling an enterprise cloud services ecosystem" , BM Journals & Magazines, 2011
[9]. W. Zeng, J. Zhao, M. Liu, J. Zhao, M. Liu, "Several public commercial clouds and open source cloud computing software", 7th International Conf. on CSE (ICCSE), . 2012
[10]. Jain, S., Choudhury, T., Kumar, P., Rathore, Y.S, Communication by terrorist using access points: Cyber security, SmartTechCon 2017.
[11]. V Choudhary et. al.,An approach to improve task scheduling in a decentralized cloud computing environment,International Journal of Computer Technology and Applications 3 (1), 312-316,2012.
[12]. Sharma, A., Choudhury, T. et. al., Health Monitoring & Management using IoT devices in a Cloud Based Framework, (ICACCE) International Conference on Advances in Computing and Communication Engineering, 2018.

Optimal tree led approach for effective decision making to mitigate mortality rates in a varied demographic dataset

Romil Choudhary
School of Computer Science and Engineering
University of Petroleum and Energy Studies
Dehradun, India.
romil.choudhary.rc@gmail.com

Monit Kapoor
School of Computer Science and Engineering
University of Petroleum and Energy Studies
Dehradun, India.
kapoor.monit@gmail.com

***Abstract*— Decision trees always have been a data structure of choice for taking decisions under various conditions and often provide elegant decision making to resolve complex conditional problems. Optimality is a condition under which parameters of interest have to either minimized or maximized. The principle of Pareto Optimality when applied to test the worth of solution nodes in a decision tree provide for a designated conditional approach to optimize the state reported in the solution nodes. In this paper, an effort has been made to test the leaf nodes of decision trees created by three different condition sets applied on a huge dataset having in excess of 106 entries with multiple attributes using Pareto Optimality principle. The results produced by Pareto Optimality Operator thus designed clearly demarcates the conditions under which the casualty rates can be minimized.**

Keywords*— *Decision Tree, Pareto Optimality, Classification Tree, Regression Tree, Model Tree

I. INTRODUCTION

Decision Tree is linked with Pareto efficiency[1] as there always exists one Pareto efficient solution to a problem there also exist a decision tree that would represent the whole domain by correctly classifying all, or at least some proportion which is significant, of its cases. While constructing a decision tree we always try to minimize error in each leaf node. There are two types of model for Decision Tree i.e. Classification trees and Regression trees.

Most of the recent research work has focus on Regression Tree [2] and also on Model Tree [3] of DT, in which the target variable can take continuous values (i.e. Real number). In the Regression tree structures, leaf nodes represent a constant value and in contrast Model Tree has this value replaced by a regression function, to predict the target values and the branches of tree represents conjunction of features that lead to those class labels. Regression Tree and model tree are not as popular as Classification Tree but they are highly competitive with different machine learning algorithms.

Classification trees are those tree models where the target variable can take a finite set of values, in these type of tree structures branches represent conjunctions of features that lead to those class labels and the leaves represent class labels. Regression tree are those decision tree where the target variable can take continuous values. Regression is also a data mining [4] task through building a model based on one or more predictors (numerical and categorical variables) for predicting the value of target variable (numerical variable).

J.R. Quinlan gave the core algorithm for building Decision Tree called ID3 which without backtracking employs a top-down search [5]. It is a greedy search through the space of possible branches and for the regression analysis the ID3 algorithm can be used to construct a decision tree by replacing information gain value with the standard deviation reduction they are based on a stepwise search procedure. Decision tree using Genetic Algorithm (GA) [6] aim is to cover the entire search space without performing an exhaustive search of the domain. GA is one of the heuristic search methods, it may take more time and space but is considered very effective in solving some problems in which brute force might not give a solution.

Decision Tree-based approach is very powerful and has been utilized to arrive at decisions in singular and multiple conditions over many years. It has been more than 50 years since the first regression tree algorithm was given [7], there have been many enhancements in Regression models for minimizing the length of the tree which give more effective solution.

II. BACKGROUND

A. Decision Tree

A decision tree is a tree, which makes a decision at each node of the tree so as to come up with a result i.e. either discrete or continuous values. It also breaks down a dataset into smaller and smaller subsets while, at the same time an associated decision tree is incrementally developed which leads closer to the solution node. It has three parts i.e.

a) Decision node:- It has 2 or more branches

b) Leaf nodes:- Represents classification or decision

c) Root node:- The topmost node which is the best predictor

978-1-5090-6785-5/18/$31.00 © 2018 by IEEE

The conventional decision tree uses the ID3 algorithm, which makes use of Entropy and Information Gain values to arrange nodes in a structure so as to minimize the depth of the tree, so that faster results could be achieved.

Decision Trees can handle both categorical and numerical data i.e. for classification or regression problem respectively. It is a very powerful tool for data mining and machine learning [8]. The main problem with this structure is that it does not have backtracking and also overfitting. For these methods like pre-pruning and post-pruning are used.

B. Regression

Regression is an analytical method to measure the association of one or more independent variables with dependent variables. In Regression technique, the output variable contains continuous or discrete values. There are three types of regression i.e. Linear, Multiple and Logistic Regression. In linear regression, a single independent variable is used to predict the value of a dependent variable by taking a linear function to represent it. In multiple linear regression, two or more independent variables are used to predict the value of a dependent variable. While in Logistic Regression the output variable takes discrete values. A line is drawn such that the sum of squared error of all the points to the line is minimized. There can also be other error measuring techniques but the sum of square error is most common [9]. Other than this we can also use a tree structure for Regression analysis in which the leaf node represents a set of continuous values. The common regression tree methods are AID and CART they are based on node impurity being the sum of squared deviations about the mean and the node predicting the sample, mean [10].

C. Pareto Optimality

Pareto optimality is a solution for multiple objectives and it is a term derived from economics.

This concept was given by Vilfredo Pareto who was Italian engineer and economist, who used the concept in his studies of economic efficiency and income distribution [11], so it is named after him as Pareto Optimality. The problems in the real world for optimization often conflict as if we improve one property then it may affect other in such a way that it could degrade their performance when one property is dependent on other property in this sort of case we can use Pareto Optimality for optimization. A Pareto optimal solution is in which any part cannot be improved without making some other part worse.

If we consider m conflicting objectives that need to be minimized. A solution A = {a1,a2,...,am} is said to dominate solution B ={b1,b2,...,bm} (symbolically denoted by A≺B) if and only if:

$$(A \prec B) \Leftrightarrow (\forall\_i)(a\_i \leq b\_i) \wedge (\exists\_i)(a\_i < b\_i) \qquad (1)$$

The solutions that are not dominated by any other solutions are constituted as Pareto optimal set of solution:

$$\{A \mid \neg(\exists B, B \prec A)\} \qquad (2)$$

The set of Pareto optimal solutions is referred as Pareto front. Pareto optimality is a popular technique in Machine learning and we are implementing it for regression task. It is an exploration of Pareto optimality in regression task [1].

III. Literature Survey

All kinds of tree representations like univariate, regression, model, oblique etc. can be evolved using Global Model Tree (GMT) [12]. Traditional univariate trees can be used to form a representation of Model trees, so every internal node is split on the basis of a unique attribute. Multiple linear regression models which are contained in the leaf nodes are created with learning instance associated with respective leaf.

The models in the leaves, tree structure and all tests within the internal nodes can be modified using GMT framework [13]. For multi-objective optimization currently there are 2 strategies implemented with the help of Global Model Trees:

1. Weight formula
2. Lexicographic Analysis.

The Bayesian Information Criterion (BIC) is one of the various weight formulas tested within the GMT system, it has the highest performance with model trees and regression. It is given by the formula:

$$Fit_{BIC}(T) = -2 * \ln(L(T)) + \ln(n) * k(T), \quad (3)$$

Where n is the number of observations in the data L(T) is the maximum of the likelihood function of the tree T, and k(T) is the number of model parameters in the tree. The L(T) function which is the log(likelihood) function is typical for regression models and can be expressed as:

$$\ln(L(T)) = -0.5n * [\ln(2\pi) + \ln\left(\frac{SSe(T)}{n}\right) + 1] \quad (4)$$

Where the sum of squared residuals of the tree T is SSe(T) [1]. The term k(T) in this measure of goodness of fit can be viewed as a penalty for over-parameterization as the tree complexity is reflected by it which for regression trees equals to the number of nodes where it is denoted by Q(T), whereas the number of attributes in the linear models in the leaves which are denoted as W(T) is also included for model trees. From SSe(T), Q(T) and W(T) one is used as the measure when the lexicographic analysis is applied in for the fitness evaluation, each pair of individuals is analyzed, in order of their priorities. The tree accuracy is prioritized first for the next number of terminal nodes so as to prevent overfitting and overgrown trees. The W(T) measure is used to keep the models in the leaves simplest and also penalizes for over-parametrization to prevent overfitting.

Pareto-Based Approach for GMT:

The set of Pareto optimal solution is found out for the goal of multi-objective optimization, which is for providing insights into the trade-o□s between the multi-objective. From

the current study, we may not get the Pareto front by GMT fitness functions: weight formula and lexicographic analysis they yield only a limited subset of the solutions.

For tackling multi-objective optimization problem multiple Evolutionary Algorithms (EAs) were developed, particularly among various dominance comparison mechanism like non-dominated sorting genetic algorithm NSG-II has been very effective in the search for a set of Pareto-optimal solutions [1]. NSG-II showed fast convergence to the Pareto-optimal set of solution. Efficient non-dominated sorting strategy (ENS) [15] is search strategy which is more recently being applied. This strategy was selected as due to its e□ciency. From some experimental evaluation, it was found that ENS had shown that it outperforms other non-dominated sorting approaches for many optimization problems which had a small number of objectives, which is here the case. Most of the non-dominated sorting methods and ENS algorithm is conceptually very di□erent. The front each solution belongs to one by one is determined by ENS, where the front of all solution is determined by typical non-dominated sorting approaches. Since a solution to be assigned is only needed to be compared with solutions that have already been assigned to the front in this way ENS avoids duplicate comparisons. The archive fronts are updated in the second step of the proposed extension.

In the elitist list all the solutions from the Pareto front are stored in it, whenever a new solution from the current population dominates one in the list, the list is updated each time. This operation is computationally more expensive but it is still not very large in case of GMT so it is acceptable as the Pareto front. The updated crowding distance procedure is adapted in proposed extension. The crowded comparison operator ($\prec_n$) helps to order (ranking) the solutions. For diversity preservation and to maintain a well-spread Pareto front in the NSGA-II the crowding-distance is used.

IV. Problem Statement and Proposed solution

The best local attribute for each internal nodes is selected in producing the optimal tree, which is performed according to some of the multi-objectives in the database. The conventional algorithms can deal only with single-objective that aggregates into multiple-objectives. Here we are going to use multiple-objective approach for Pareto efficiency. The dataset we are using is a multi-relational dataset, which is about potential excess deaths. Given attributes like locality, the cause of death, year, age, state etc. what is the excess death i.e. (observed deaths - expected deaths) under following conditions?

We will focus on four attributes as provided next:

1. Age: age<=49 or age>49.
2. Year: 2005-2010 or 2010-2015.
3. Locality: metropolitan or nonmetropolitan.
4. Cause of Death: heart disease, cancer, stroke or other.

From these attributes, we find a conditional set, which dominates over all other so that excess death decreases and it could help implement methodologies to take place to prevent deaths of people in different environments. This dominating conditional set is being found by using Pareto optimality condition.

We are going to use a special type of Decision Tree whose leaf node will contain an array containing observed death - expected death i.e. excess death. As it is having multiple elements at the end of node it is not a simple decision tree. According to the decision parameters like age group, the cause of death, locality, and year various decision would be made and the array will be filled and it would be having data of various states and regions getting filled at every element of the array. There is a condition set C $\{C_1, C_2, C_3\}$ on which we will be working.

Notation Table:-

Notation	Sub Notation	Description
A_p	A_{p1} A_{p2}	Age<=49 Age>49
L_p	L_{p1} L_{p2}	Metropolitan area Nonmetropolitan area
Y_p	Y_{p1} Y_{p2}	2005<year<=2010 2011<year<=2015
D_p	D_{p1} D_{p2} D_{p3} D_{p4}	Heart Disease Cancer Stroke Other
E_p	-	Excess death in the population.

C_1:- (A_p && L_p && D_P)
C_2:- (A_p && Y_p && D_p)
C_3:- (L_p && Y_p && D_p)
Next, we will find the most dominant condition i.e. $C_d \prec C_i$ (i=0,1 and 2).

V. Experiment and Results:

Dataset available at [16] is being used to get a Pareto optimal tree that has to be generated using conditional set C_1(A_p && L_p && D_P), C_2(A_p && Y_p && D_p) and C_3(L_p && Y_p && D_p), so that efficient planning can be done to mitigate excess deaths. The root node and decision nodes of the tree are used to make decisions and all the leaf nodes contain an array which has observed death - expected death values which need to be minimized. There will be 3 trees from the conditional set C_1, C_2, and C_3, comparing the leaf nodes which contains an array of excess deaths, with other tree's array we get the optimal tree i.e. the tree which has less number of deaths. We rank each tree based on this and see which one is more Pareto

optimal. We have dropped the extra leaf nodes from the tree if exists for simplification purpose. Pareto optimality is used for this type of multi-relational dataset as it takes into consideration all the attributes that provide a single optimal result.

Pseudocode:-

```
program decisionTree
  Open "cod.txt"
  for each line
     set i=0;
     split line by ","
       column(i) =word
       i= i+1
     set year=columns(0), age=columns(5), ed=columns(11)
     set  locality=columns(7), cod=columns(1)
     set tree1 root1= agep, root2= locality, root3= cod
     set tree2 root1= age, root2= year, root3= cod
     set tree3 root1= locality, root2= year, root3= cod
     set  arr1=(Ap  && Lp && DP), arr2=(Ap  && Yp && DP)
     set arr3=(Lp  && Yp && DP)
  end for
  for i=0 to len(arr1)
     if arr1<arr3
        p1=p1+1
     else
        p3=p3+1
     end if
     i=i+1
  end for
  if p1>p3
     percent=(p1*100)/(p1+p3)
      v=1
      print "C1 beats C3 by "  percent "%"
      p1=0
      l=1
      for i=0 to len(arr2)
         if arr1<arr2
            p1=p1+1
         else
            p2=p2+1
         end if
         i=i+1
      end for
  else
      percent=(p3*100)/(p1+p3)
      v=3
      print "C3 beats C1 by "  percent "%"
      p3=0
      l=3
      for i=0 to len(arr2)
         if arr3<arr2
            p3=p3+1
         else
            p2=p2+1
         end if
         i=i+1
      end for
  end if
   if l<p2
      l=2
      if v=1
         percent=p2/(p1+p2)
      else
         percent=p2/(p2+p3)
      end if
      percent=percent*100
      print "C2 beats Cv by "  percent "%"
   else
      if v=1
         percent=p1/(p1+p2)
      else
         percent=p3/(p2+p3)
      end if
      percent=percent*100
      print "Cv beats C2 by "  percent "%"
   end if
end
```

We have implemented the proposed algorithm as provided in pseudocode using C++ Language. Comparisons were made by comparing different arrays created by the tree structure to test the Pareto Optimality condition as provided in [1] where aggregated leaf nodes were compared with one another. The leaf nodes have been generated by virtue of three decision trees generated by three conditionals provided in Set C above. On implementing the model we found that C_3 dominates C_1 by 58% and C_2 dominates C_3 by 56%.This gives confidence that C_2 condition which is an aggregation of Age, Year and Cause of Death will lead to lesser deaths by 56% under condition C_3. Progressively C_3 is a better condition set that C_1 by 58%, which gives us a confidence that C_2 condition would reduce casualty rates by more than 84% holistically if taken up as a policy decision to curtail death risks.

Therefore, C_2 is optimal conditional set according to Pareto optimality to minimize the number of deaths and hence policymakers can take a decision made on the basis of condition C2 to the demography under discussion as reported in dataset provided at [16].

VI. Conclusion and Future Scope

Pareto optimality principle is found to be highly suitable for a multi relational dataset, as one attribute can't be improved without adversely impacting other. So, the solution which we get is optimal and it can be easily performed on large datasets to find a single optimal solution to a multi-relational dataset with higher efficiency and good accuracy.

The purpose of this study was to establish the conditions under which the casualty rates expected v/s the rates observed need to be identified on basis of varying demographic parameters. The demographic parameters in the dataset used in this work at [16] incorporate many attributes like urban v/s rural, different diseases and varying age groups. A simple decision tree would have complex conditions to evaluate but Pareto optimality allows us to crisply identify the aggregated condition set under which death rates are found to be minimal. This aggregated condition can help the policymakers to adapt to the aggregated conditions and adapt to those aggregated conditions to minimize the casualty rates.

Once a node is selected, we cannot backtrack and it is a potential weakness of this algorithm as it behaves like a decision tree, which just takes in account, which steps need to be taken at a particular node not using backtracking. In our

978-1-5090-6785-5/18/$31.00 © 2018 by IEEE

future works, we shall try to reduce the effects of this drawback by introducing the principle of backtracking to the Pareto optimal trees generated in this research work by using neural networks.

References

[1] Marcin Czajkowski and Marek Kretowski "A Multi-objective Evolutionary Approach to Pareto Optimal Model Trees. A Preliminary Study" (2016)

[2] Breiman, L., Friedman, J., Olshen, R., Stone, C.: Classification and Regression Trees. Wadsworth and Brooks, Monterey (1984)

[3] Barros, R.C., Ruiz, D.D., Basgalupp, M.P. " Evolutionary model trees for handling continuous classes in machine learning". Inf. Sci. 181(5), 954–971 (2011)

[4] Rokach, L., Maimon, O. "Data Mining with Decision Trees: Theory and Applications". World Scientific Publishing Co. Inc., River Edge (2008)

[5] Quinlan, J. R. 1986. "Induction of Decision Trees. Mach. Learn". 1, 1 (Mar. 1986), 81–106

[6] Mitchell Melanie, "An Introduction to Genetic Algorithms". (1998)

[7] Loh, W.Y. "Fifty years of classification and regression trees". (2014)

[8] Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R. "Advances in Knowledge Discovery and Data Mining". American Association for Artificial Intelligence, Menlo Park (1996)

[9] Bishop, "Pattern Recognition and Machine Learning" (2006)

[10] Wei-Yin Loh,"Classification and regression trees". (2011)

[11] Charles J. Petrie, Teresa A. Webster, Mark R. Cutkosky "Using Pareto Optimality to Coordinate Distributed Agents" (1995)

[12] Czajkowski, M., Kretowski, M. "The role of decision tree representation in regression problems - an evolutionary perspective" (2016)

[13] Czajkowski, M., Kretowski, M. "Evolutionary induction of global model trees with specialized operators and memetic extensions". (2014)

[14] Deb, K., Pratap, A., Agarwal, S., Meyarivan, T. "A fast and elitist multiobjective genetic algorithm: NSGA-II". IEEE Trans. Evol. Comp. 6(2), 182–197 (2002)

[15] Zhang, X., Tian, Y., Cheng, R., Jin, Y. "An e□cient approach to nondominated sorting for evolutionary multiobjective optimization". IEEE Trans. Evol. Comput. 19(2), 201–213 (2015)

[16] https://data.cdc.gov/NCHS/NCHS-Potentially-Excess-Deaths-from-the-Five-Leadi/vdpk-qzpr [Accessed on September 2017]

Algorithm implemented can be found at https://github.com/RomilChoudhary1/Optimal-Decision-making-using-Pareto-Optimality

978-1-5090-6785-5/18/$31.00 © 2018 by IEEE

Breast Cancer Detection Using Machine Learning Algorithms

[1]Shubham Sharma, [2]Archit Aggarwal, [3]Tanupriya Choudhury
[1]lftshubhamsharma@gmail.com,[2]archit.aggarwal1508@gmail.com, [3]tanupriya1986@gmail.com
[1,3] University of Petroleum & Energy Studies (UPES),Dept. of Informatics, School of Computer Science, Dehradun
[2] Amity University Uttar Pradesh

Abstract: The most frequently occuring cancer among Indian women is breast cancer. There is a chance of fifty percent for fatality in a case as one of two women diagnosed with breast cancer die in the cases of Indian women[1]. This paper aims to present comparison of the largely popular machine learning algorithms and techniques commonly used for breast cancer prediction, namely Random Forest, kNN (k-Nearest-Neighbor) and Naïve Bayes. The Wisconsin Diagnosis Breast Cancer data set was used as a training set to compare the performance of the various machine learning techniques in terms of key parameters such as accuracy, and precision. The results obtained are very competitive and can be used for detection and treatment.

***Keywords*— Breast Cancer, random forest, k-Nearest-Neighbor,**
naive bayes

I. INTRODUCTION

The most commonly occurring type of cancer is breast cancer. It is known to affect over two million women annually. For women diagnosed during 2010-14, five-year survival for breast cancer shows very heavy variation with changes in location. It is generally known to be above fifty 50% in most places. There are no prevention techniques for breast cancer but early detection and diagnosis is critical in determining the chances of survival.

During the early stages of the disease, the symptoms are not presented well and hence diagnosis is delayed. It is recommended by the NBCF (National Breast Cancer Foundation) that women over the age of forty years of age should get a mammogram once a year. A mammogram is an X-ray of the breast. It is a medical technique used for the detection of breast cancer in women without any side effects deeming the procedure as safe. Women who get regular mammograms have a higher survival rate as compared to women who do not. According to [2] in 2018, over six hundred thousand fatalities were caused by breast cancer. The number is approximately fifteen percent of the total deaths resulting from all types of cancer among women. The chances of contracting this particular type of cancer are usually higher in urban regions, however, the rate of contraction seems to be on an upward rising trend globally. The only current method of improving the results of breast cancer cases is early diagnosis and screening.

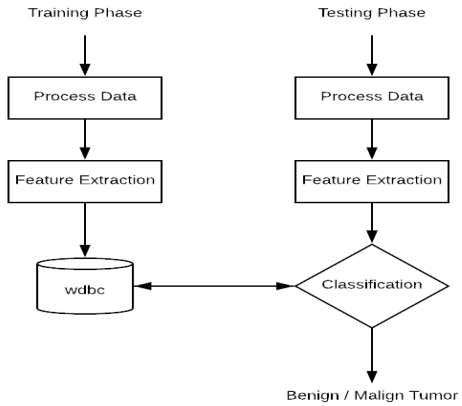


Figure 1: Proposed Breast Cancer Detection Model

II. RELATED WORK IN BREAST CANCER

Breast cancer detection using Relevance Vector Machine [3], obtained an accuracy of 97% using Wisconsin original dataset which has 699 instances and 11 attributes, while [4] allots distinct weights to different attributes with regard to their capabilities of prediction and yielded an accuracy of 92% working with the weighted naïve bayes method. [5] built a hybrid classifier of Support Vector Machines and decision trees in WEKA and obtained an accuracy of 91%. [6] used Linear Discriminant Analysis for feature selection and trained the dataset by using one of the fuzzy inference method called Mamdani Fuzzy inference model and obtained an accuracy of 93%.

Various differentiation between multiple techniques has been provided through this manuscript[7] like Bayes Network, Pruned Tree, kNN algorithm using WEKA on breast cancer dataset, it has a total of 6291 data and a dimension of 699 rows and 9 columns. The highest accuracy is 89.71% which belongs to bayes network.[11][12][13]

III. MACHINE LEARNING ALGORITHMS

Machine learning(ML) may be defined as a subset of Artificial Intelligence that inculcates the ability of learning into a system on the basis of a data set used for the purpose of training in contrast to the normal approach of coding all

1

possible outcomes before hand. Multiple approaches and techniques are present to making systems which can learn. Some of them are neural networks, decision trees and clustering.

A. ML is to be broadly categorised under three categories namely - reinforcement learning, supervised learning and unsupervised learning and.

1) Supervised Learning: generates a function predicting outputs based on input observations. The function is generated from the training data and guides the system to produce useful epiphanies for new data sets introduced to the system.

2) Unsupervised Learning: Learning In this technique, the machine is forced to train from an unlabeled dataset and then differentiating it on the basis of some characters and allowing the algorithm to act on that information without external guidance.

3) Reinforcement Learning: The learning process continues from the environment in an iterative fashion. All possible system states are eventually learned by the system over a prolonged period of time.

B. Random Forest

It is a *supervised learning* algorithm. An ensemble of decision trees is created, the bagging method is used to train the system.

The ground methodology on which this technique is based is recursion. A random sample of size N is picked from the data set in each instance of an iteration.

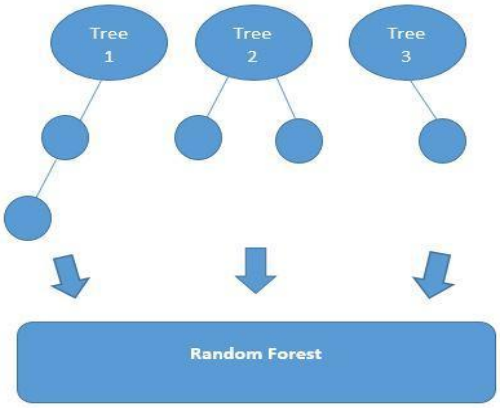


Figure 2: How Random Forest Works

The dataset has been divided into training and testing sets, there are 398 observations for training set and 171 observations for testing. The number of estimators are set to 72 thus it is ensured that every observation is predicted at least a few times. It is obvious that diagnosis, radius_mean, texture_mean, perimeter_mean are influential variables, the other variables are of moderate influence but none of them can be neglected to increase the model accuracy.

The confusion matrix of random forest is quite promising. There are only five observations that are misclassified as Benign and four observations are misclassified as Malignant and the accuracy equals 94.74%.

		Predicted	
		Benign	Malignant
Actual	Benign	103	5
	Malignant	4	59

Table 1: Random Forest Confusion Matrix

C. K-Nearest-Neighbor (kNN)

K may be seen as the representation of the data points for training in close proximity to the test data point which we are going to use to find the class. A k-nearest-neighbor may be defined as the algorithm used to determine where a data set belongs to on the basis of the other data sets present around it. The technique is a supervised learning approach used for regression and classification. To process a new data point, KNN gathers all the data points close by to it. Attributes which have a large degree of variation are key factors in determining the distance.

Given N training vectors in the Figure 3, kNN algorithm identifies the k nearest neighbors of regardless of labels.

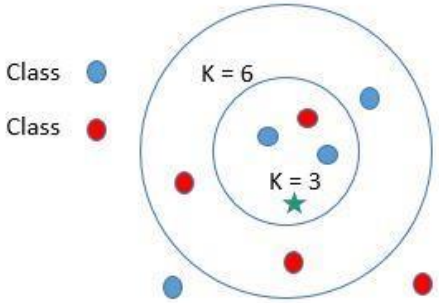


Figure 3: kNN Illustration

The accuracy of kNN is found to be 95.90% , there is only one observation that is misclassified as Benign and four observations are misclassified as Malignant as represented in Table 2. The results are comparatively better than Random Forest algorithm.

		Predicted	
		Benign	Malignant
Actual	Benign	107	1
	Malignant	6	57

Table 2: kNN Confusion Matrix

D. *Naïve Bayes*

Classifiers which are probabilistic in nature, based on the application of Bayes theorem may be defined as Naive Bayes classifiers. It is naïve because it assumes that all features are independent from each other, this is generally not the case in real life scenarios, but still Naïve Bayes proves to be efficient for wide variety of machine learning problems.

There are sixteen misclassified observations, seven of them being benign and nine of them are malignant.

The same 398 observations are used for training set and 171 observations for testing and the accuracy equals to 94.47%.

		Predicted	
		Benign	Malignant
Actual	Benign	101	7
	Malignant	9	54

Table 3: Naïve Bayes Confusion Matrix

E. *Comparison Among Proposed Algorithms*

Each one of the three algorithm's – kNN, Naïve Bayes and Random Forest have their advantage and disadvantage over each other in terms of performance, the type of problem they handle etc. As shown in Table 4: kNN test time is O(1) without preprocessing of training set [8], in the case of Naïve Bayes: N is the number of training examples and d is the dimensionality of the features whereas for Random Forest [9]: N is the number of samples and K is the number of variables randomly drawn at each node. Naïve Bayes algorithm deal only with classification problems whereas both kNN and Random Forest can deal with classification as well as regression problems. In terms of accuracy both kNN and Random Forest can deliver high accuracy but Naïve Bayes algorithm need large number of records in order to yield a better accuracy. Algorithms that simplify the function to a known form are called parametric machine learning algorithms, Naïve Bayes algorithm can be expressed as parametric as well as non-parametric model.

Parameter	KNN	Naïve Bayes	Random Forest
Time Complexity (Training Phase)	O(1)	O(Nd)	$\Theta(MKN\log^2 N)$
Problem Type	Classification & Regression	Classification	Classification & Regression
Accuracy	Provides high accuracy	For high accuracy it needs very large number of records	Provides high accuracy
Model Parameter	Non Parametric	Parametric/Non Parametric	Non Parametric

Table 4: Comparison among kNN, Naïve Bayes and Random Forest

IV. PROPOSED METHODOLOGY

A. *Dataset Description*

The project is based on Wisconsin Diagnosis Breast Cancer data set.The data set has been obtained from the 'UCI ML' repo, it has 569 instances and 32 attributes and there are no missing values. The output variable is either benign (357 observations) or malignant (212 observations). The most influential variables are diagnosis, radius_mean, texture_mean, perimeter_mean, area_mean etc. The positive class is used to for benin cases and the negative class is used in malignant cases. The k-fold cross-validation is utilised in which the presented data is divided into k eqully sized bits.

Dataset	**No. of Attributes**	**No. of Instances**	**No. of Classes**
Wisconsin Diagnosis Breast Cancer(WDBC)	32	569	2

Table 5: Description of WDBC Dataset

B. *Performance Metrics*

This sections describes the parameters that are used for measuring performance of machine learning techniques.

A confusion matrix for actual and predicted class is derived comprising of the standard five values namely TruePositive, FalsePositive, TrueNegative and FalseNegative to evaluate the performance.

1. *Accuracy*

Accuracy is a good predictor for the degree of correctness in the training of the model and how it may perform generally. It may be defined as the measure of the correct prediction in correspondence to the wrong ones. Thus the equation presented can be used to calculate the value of accuracy:

$$Accuracy = \frac{(TruePositive + TrueNegetive)}{(\text{TruePositive} + \text{FalsePositive} + \text{TrueNegative} + \text{False Negative})}$$

2. *Recall*

Recall known as sensitivity in general terms, may be defined as the ratio of rightfully determined positive instances to the all observations. Recall may be seen as a measure for the effectiveness of the system in predicting positives and determining costs.

$$Recall = \frac{TruePositive}{(TruePositive + FalseNegetive)}$$

3. *Precision*

The degree of correctness in determining the positive outcomes may be defined as precision. It is basically the ratio between true positives and the overall set of positives. This depicts the handling capacity of the system for positive values but does not provide insight into the negative values.

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

4. *F1 Score*

It is the weighted average of Precision and Recall. This measure hence, considers both type of false values. F1 score is considered perfect when at 1 and is a total failure when at 0.

$$F1\ Score = \frac{2*(Precision*Recall)}{(Precision + Recall)}$$

V. IMPLEMENTATION AND RESULT ANALYSIS

A comparative study using Random Forest, kNN (k-Nearest-Neighbor) and Naïve Bayes algorithm which are implemented in a computer having configuration as Intel Core i7 with 16GigaBits RAM has been proposed. We have used numpy, pandas and Scikit-learn which are open source machine learning libraries in Python. An open source web application named as Jupyter Notebook is used to run the program.

The classifier was tested using the k − fold cross validation method. We have utilized the 10 fold technique that is the data set segregated in ten different chunks. Nine out of the folds used in the system are used for training and the last set is used for the purposes of testing and analysis. We have utilized 398 observations for training set and 171 observations for testing out of 569 observations. The graphical representation of the performance metrics for the three illustrated algorithms are shown in Figure 4. The results presented in Table 6 shows that Random Forest's has the best *recall* performance measure but kNN has the best *accuracy*, *precision* and *F1 Score* over Naïve Bayes and Random Forest.

Model Performance (Testing Phase)			
	RF	kNN	Naïve Bayes
Accuracy (%)	94.74	95.90	94.47
Precision (%)	92.18	98.27	88.52
Recall (%)	93.65	90.47	85.71
F1 Score (%)	92.90	94.20	87.09

Table 6: Performance Measure Indices

Figure 4: Graphical representation of Performance Measure Indices

VI. CONCLUSION

The most frequently occurring type of across cancer is breast cancer. There is a chance of twelve percent for a women picked randomly to be diagnosed with the disease[10]. Thus, early detection of breast cancer can save a lot of valuable life. The proposed model in this paper presents a comparative study of different machine learning algorithms, for the detection of breast cancer. Performance comparison of the machine learning algorithms techniques has been carried out using the Wisconsin Diagnosis Breast

Cancer data set. It has been observed that each of the algorithm had an accuracy of more than 94%, to determine benign tumor or malignant tumor. From Table 6, it is found that kNN is the most effective in detection of the breast cancer as it had the best accuracy, precision and F1 score over the other algorithms.

Thus supervised machine learning techniques will be very supportive in early diagnosis and prognosis of a cancer type in cancer research.

REFERENCES

[1] National Institute of Cancer Prevention and Research, cancer statistics [Online], Available: http://cancerindia.org.in/statistics/

[2] WHO breast cancer statistics [Online]. Available: http://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en/

[3] B.M. Gayathri, Dr. C.P. Sumathi, "Comparative study of relevance vector machine with various machine learning techniques used for detecting breast cancer" 2016

[4] S Kharya and S Soni,"Weighted Naïve Bayes classifier –Predictive model for breast cancer detection", January 2016

[5] Sivakami, "Mining Big Data: Breast Cancer Prediction using DT-SVM Hybrid Model" 2015

[6] B.M.Gayathri and C.P.Sumathi,"Mamdani fuzzy inference system for breast cancer risk detection", 2015.

[7] Mohd,F.,Thomas,M, "Comparison of different classification techniques using WEKA for Breast cancer" 2007.

[8] Time complexity and optimality of kNN [Online] Available: https://nlp.stanford.edu/IR-book/html/htmledition/time-complexity-and-optimality-of-knn-1.html

[9] Gilles Louppe, "Understanding Random Forests from theory to practice" 2015.

[10] U.S. Breast Cancer Statistics [Online] Available: https://www.breastcancer.org/symptoms/understand_bc/statistics

[11] T Choudhury, V Kumar, D Nigam ,An Innovative Smart Soft Computing Methodology towards Disease (Cancer, Heart Disease, Arthritis) Detection in an Earlier Stage and in a Smarter Way-International Journal of Computer Science and Mobile Communication (IJCSMC) 2014.

[12] T Choudhury, V Kumar, D Nigam, B Mandal ,Intelligent classification of lung & oral cancer through diverse data mining algorithms, International Conference on Micro-Electronics and Telecommunication Engineering 2016

[13] T Choudhury, V Kumar, D Nigam,Intelligent Classification & Clustering Of Lung & Oral Cancer through Decision Tree & Genetic Algorithm - International Journal of Advanced Research in Computer Science and Software Engineering, 2015

Salient Features of an Effective Immersive Non-Collaborative Virtual Reality Learning Environment

Rhodora Abadia
Torrens University Australia
Adelaide, South Australia
rabadia@laureate.net.au

James Calvert
Torrens University Australia
Adelaide, South Australia
jcalvert@laureate.net.au

Syed Mohammad Tauseef
University of Petroleum and Energy Studies
Dehradun, India
smtauseef@ddn.upes.ac.in

ABSTRACT
The use of immersive virtual reality learning environments (VRLEs) is changing the way students learn and understand things. A VRLE allows its users to get immersed in the simulation, thus giving the sense of being part of the real world that it represents. Virtual Reality (VR) existed in various forms in the past two decades, but its early adoption in education was hampered by its high cost. Emergence of affordable head-mounted displays (HMD) is now making it possible to provide VR experience in classrooms. This paper aims to apply integrative review of relevant studies conducted in evaluating the effectiveness of VRLEs and in doing so reveal the existing research gaps among VRLE studies. This paper identifies salient features of a fully immersive and non-collaborative VRLEs that uses HMD. Salient features were derived from evaluating different instruments used to measure the effectiveness of immersive VRLEs. A framework and example of instruments to evaluate salient features of an effective VR, using Kokoda VR as a case study, are also provided.

CCS Concepts
• **Applied Computing →Education →Interactive Learning Environments**

Keywords
Virtual Reality Learning Environment; Immersive Learning Environment, Technology in Education; Salient Features

1. INTRODUCTION
For many years, schools and universities had to change to way they teach in order to employ the latest technology to provide the best learning experience. From the use of software like PowerPoint, to the introduction of multimedia and online learning technologies. The introduction of new technologies being designed and used specifically for education is changing the way students learn and understand things. One such technology is virtual reality (VR).

VR has existed in various forms in the past two decades, but high cost became a main barrier for its early adoption in education, outside of experimental studies [11]. The emergence of affordable head-mounted displays (HMDs) is now making it possible to provide immersive virtual reality experience in classrooms. Virtual reality learning environments (VRLEs) is an application of VR in education which allow its users to get immersed in the simulation, thus giving the sense of being part of the real world that it represents.

Immersive VR is considered the next frontier for education. It is a powerful platform that gives students the opportunity to interact with content in three-dimension thereby providing a personal and meaningful experience. Although immersive VR seems to offer promising educational benefits, there are still many issues that need further investigation. As with any new technology applied in education, it is important to note that this technology is merely a tool that need to be carefully and effectively designed and implemented [20]. Salient features need to be identified before designing and implementing such technology.

1.1 Case Study
In July 2017, Torrens University Australia and the Australian Broadcasting Corporation (ABC) collaborated to develop Kokoda VR, a fully interactive Virtual Reality experience that immerses students in the main events of the Kokoda Track campaign in WWII. Kokoda VR is a pilot for an innovative production methodology which incorporates the photogrammetry method of scanning real objects and locations to be used in the virtual world [1]. It also features original museum artefacts, historical interviews and videos. Unlike 360˚ video, a 3D reconstruction of a location using photogrammetry allows users in VR to walk around the scene, pick up and inspect objects. Soundscapes that respond to the user's location and self-guided actions further increase the immersion in the experience. Autonomous exploration of these real-world locations increases the agency of the user.

The Kokoda VR's purpose was for it to be used in classrooms for high school students and older; and be used as supplement when discussing Australian WWII history. However, there is still a need to study if Kokoda VR is an effective learning tool. Do learners construct meanings they engage with the virtual reality experience? There are issues that needs to be studied further on the effectiveness of using immersive, realistic and interactive virtual worlds as tools in education. To be able to evaluate the effectiveness of VRLEs like Kokoda VR, the first step is to identify salient features specific to the use of immersive VR in education.

This paper aims to evaluate relevant studies conducted in analyzing the effectiveness of VRLEs and by doing so reveal the existing research gap among VRLE studies. This paper proposes salient features of an effective immersive VRLEs such as those

ICETC 2018, October 26–28, 2018, Tokyo, Japan

ACM ISBN 978-1-4503-6517-8/18/10…$15.00
https://doi.org/10.1145/3290511.3290558

achievable through the use of HMDs and that are non-collaborative (or individualized learning experience). These salient features focus on VRLEs that are fully immersive and were derived from evaluating different measures used to assess the effectiveness of an immersive VRLEs.

The next section presents the approach used to review these studies and identify the salient features. It will then be followed by the presentation of the results. Examples of instruments in assessing these features using Kokoda VR as case study are also provided.

2. APPROACH

In the current analysis of studies in evaluating effective VRLEs, we use the integrative review approach that involve summarizing the data achieved by collating data [39] .

2.1 Data Source and Search Strategies

The following strategies were employed to identify studies to include in the review:

1. Electronic searches were performed using the Web of Science, one of the highly recognized databases indexing essential journals, to search for research articles.
2. Keyword search terms include, virtual reality, immersive learning environment, virtual learning environment.
3. The next step was to refine it to search words such as education, learning, and evaluation.

2.2 Inclusion and Exclusion Criteria

Studies were either included or excluded based on the following criteria:

1. The time span was from 2013-2018 to focus on the current status of VRLEs and the period when high quality HMDs became more accessible.
2. Studies that focus on just the use of virtual reality (not augmented reality and mixed-reality, or combination of these technologies).
3. Studies that used non-collaborative use of virtual reality.
4. Studies that measured the effectiveness of the VRLEs.
5. Studies that used head mounted displays (HMDs).
6. Studies that were developed specifically for learners with disabilities were excluded from the study.

2.3 Analysis

The articles were analyzed using an integrative review method [42] which includes data reduction, comparison and interpretation. In data reduction, the data in each article was categorized in a matrix according to the type of VRLE, objectives, evaluation methods, and features measured (see example in Table 1).

Table 1: Example data reduction.

Reference	Application	Objective	Evaluation Method	Salient Feature
Bharathi et al. [6]	HMDs used in interactive VLE for online engineering design activities	Test if immersive VRLE improve task performance	Task-based performance evaluation (completion)	Performance

The next step was to make comparison across the different studies guided by the aim of the research. Salient features were identified by finding patterns and clustering features by their relatedness to each other. Lastly, the results of the review were synthesized and interpreted in relation to different studies.

3. RESULTS

3.1 Study Sample

An initial search returned 13,453 hits that match the keywords and time span criteria. After refining the search using the words education, learning and evaluation, 169 articles were considered for full review. Descriptive analysis was used on the studies included and the first author used the inclusion and exclusion criteria as the guide. Another author was asked to independently asses for confirmation. The reviewers disagreed on 2 studies, yielding a 92.3% agreement and a Kappa coefficient of 0.98. The two papers were discussed, and a consensus was reached. Articles excluded either involves collaborative environment (n=19); specific for people with disabilities (n=10); used mixed-reality(n=24); or used non-HMD VR, such as desktop or projected screen (CAVE) (n=92). A total of 24 studies qualified to be included in the study.

The 24 articles reported findings from different applications of VR in education but mostly in science and engineering activities (n=8, followed by safety training (n=4); experiential (place and nature) (n=4); space education (n=3); medical(n=2); music (n=1); narrative stories (n=1); and construction education (n=1).

Comparing the features measured in the studies, features that exemplifies the same thing or goal is coded in the same name and identified as a salient feature. For example, understanding of concepts and test learning of the concepts were all coded under mastery learning. There were 12 salient features identified in the sample studies. These are: presence, mastery learning, perception, immersion, performance, engagement, embodiment, usability, knowledge retention, empathy, motivation, elicit emotion. Empathy and elicit emotion were not coded as the same feature because the intent used in VRLE evaluations were different. Eliciting emotion focuses on the user's emotional state (e.g., joy, anger and anxiety) while the goal of empathy is to measure the user's ability to understand someone's emotion. The same with mastery learning and knowledge retention. They were coded into separate features because mastery learning measures student's mastery on unit tests before moving on to next [8], while knowledge retention is defined as the quantity of knowledge retained by an individual after a specific interval of time [13].

Scrutinizing the features measured by the samples revealed that presence is the most measured feature in the studies having 37.5% (9/24). Figure 1 shows the details of the proportion of the features. It also revealed that majority of the samples (54.2%; 13/24) carried out multi-feature evaluations of VRLEs.

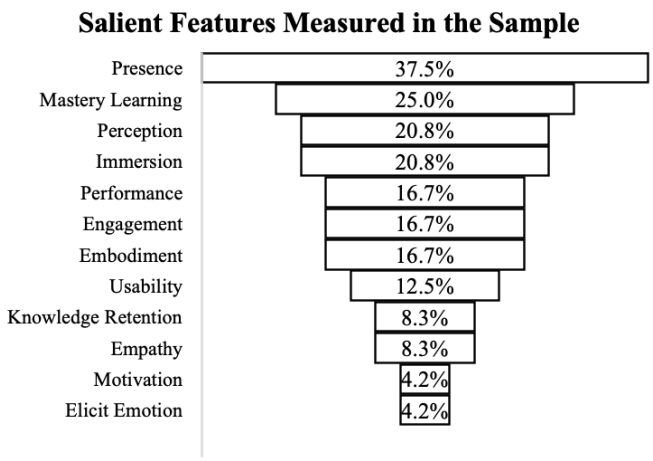


Figure 1: Proportion of the features measured in the sample.

Assessing the publication year of the samples, it shows that there is a significant increase of evaluations of VRLEs using head-mounted displays in the past two years (2017-2018). This can be attributed to the increasing affordability and availability of VR devices.

3.2 Salient Features of Immersive and Non-collaborative VRLEs

The 12 salient features identified for immersive and non-collaborative VRLEs are presence, immersion, perception, mastery learning, knowledge retention, performance, motivation, engagement, usability, embodiment, empathy, and elicit experience.

These salient features were categorized into three classifications: technical, human factors and learning. Table 2 shows the classification of salient features. The technical classification includes features that are mainly triggered by the VR technology being used (e.g., VR devices); human factors include features that are influenced by understanding psychosocial factors such as how humans perceive, engage and are affected by stimuli; and learning include features that are used to measure the students learning.

Table 2: Classification of Salient Features

	Technical	Learning	Human Factors
SALIENT FEATURES	Usability	Knowledge Retention	Presence
	Immersion	Mastery Learning	Perception
	Embodiment	Motivation	Engagement
		Performance	Elicit Emotion
			Empathy

Recent studies have shown that the VR industry‘s development focus in the next coming years is in perception, interaction and content creation [16]. Bamodu & Ye [5] has the same view stating that the quality of any VR is determined by focusing on features such as interaction and perception but adding immersion as another important feature.

3.2.1 Presence, Immersion and Perception

The most measured feature is presence having 37.5% (9/24) of the sample using it (e.g., [9,15,21,22,23, 27,28 34,35]). Effectiveness of virtual environments is often linked to the sense of presence. Presence is important in virtual world learning where it represents how close the interactions and presentations mimic the real-world experiences [40].

The most common features that were measured together were presence and immersion, even with immersion measured by only 20.8% (5/24) of the samples [22,24,28,34,35]. Samples that measured immersion also measured presence.

Immersion and presence are two interrelated but different features of VR that are often used to evaluate effectiveness of VRs. Immersion is a “technology-oriented” aspect of VR that gives the brain the impression that we’re in another place through visual information, audio or haptic feedback. The more that a system delivers displays in all sensory modalities, the more it is “immersive” [26]. Presence is the psychological, perceptual and cognitive consequence of immersion [36]. Slater et al. [36] has distinguished immersion from presence, with the former indicating a physical characteristic of the medium itself for the different senses involved. The sense of presence is indeed the subjective sense of being in the virtual environment. It is said to be the human reaction to immersion. It is the psychological perception of "being in" or "existing in" the VE in which one is immersed. Sense of presence seems to be related to immersion and several studies have shown that immersive capabilities of technology, specifically VR technology, impact the subjective feeling of presence [36].

All samples that measured presence and immersion used self-reported questionnaires as their evaluation method. Some samples included physiological measures such as the use of electro-dermal activity (EDA) [27,35]; blood volume pulse (BVP) [27]; and simulator sickness questionnaire[28]. Shin et al. [35] and Shin [34] included in their evaluation methods focus groups and interviews. Dang et al. [22] used immersive tendencies and presence questionnaires to measure subjective presence using different methods and one of them is VR observation in clinical training. The use of VR here is for purely observation and does not allow students to interact and pick-up objects. This study acknowledged that learning performance should have also been measured.

Considering that all samples used HMDs, the number of samples that measured immersion is small. Either the studies assume that immersion is there by default because of the technology used or it was removed out of the study and focus was on the VRLE. The incompleteness of the evaluation is already seen in these cases.

Presence and perception are also interrelated. Our perception is what we base our decisions on and mostly determines our sense of presence in the environment [19]. Perception in VR is the term used to describe how we take information from the virtual world and build understanding from it. Perception in this study was measured by 20.8% of the sample [29, 30,32, 33,35]. Perception is often measured by asking students if they think that using VRLEs is better compared their traditional way of learning, such as classroom or laboratory [32]. Despite the relationship between presence and perception, only Shin et al. [35] have evaluated these features together.

3.2.2 Mastery Learning, Retention, Performance and Motivation

Some VRLEs focused on measuring its effectiveness through students learning. There are four different aspects of learning that were measured: mastery learning, knowledge retention, performance (or task-based learning) and motivation.

Mastery learning, which was introduced by Bloom [8], maintains that students must achieve a level of mastery (e.g., 90% score on a test) in pre-requisite knowledge before moving forward to learn subsequent information. Mastery learning is the second highest feature (25%; 6/24) that were measured by the samples [7,13,16,24,29,30]. Most of the samples used pre and post-tests to measure learning [7,18,24,29,30] while Butt et al. [14] wanted to identify if using VR promotes mastery learning and asked participants to answer two questions regarding the topic of study. Instead of measuring exact learning, Kuronen-Stewart et al [29] focused on the overall experience using quantitative questionnaires to measure perceived effectiveness in learning and similarity to real-life setting.

Knowledge retention is defined as the quantity of knowledge retained by an individual after a specific interval of time [12]. Only two of the samples measured knowledge retention (8.3%). Butt et al.[14] tested both the mastery learning and knowledge retention by conducting test two weeks after the students participated in the study. Butussi and Chittaro [16] tested for knowledge retention before, immediately after and one week after the study.

Performance was measured in some samples using task-based activities. Task-based approach assumes learning revolves around the completion of meaningful tasks. Bharathi & Tucker [6] used the time in completing task in measuring performance. Pirker et al [32] used questionnaires and recorded quotes from students when performing the tasks while Bhargava et al. [7] used post-cognition questionnaires.

Motivation directs behavior toward particular goals and increases initiation of and persistence in activities [31]. This feature is closely related to task-based activities. However, only Pirker et al [32] measured both performance and motivation.

3.2.3 Engagement

Engagement in VRLEs is a feature that is often measured with learning and training effectiveness. But before any learning can occur, users must be sufficiently engaged. Prior researches show that engagement influences learning [2,25,44]. 16.7% (n=4) of the studies measured engagement [2,15,16,32]. Pirker et al. [32] used questionnaires and quotes while conducting the study to measure both engagement and performance. Buttussi et al.[16] used questionnaires to identify engagement and learning. Their findings include that emotional arousal is beneficial in the retention of learned concepts, retention of knowledge was higher for those who use immersive VR and that immersive VR induced higher emotional response.

3.2.4 Empathy, Embodiment, and Elicit Emotion

Empathy can occur when a VR user is immersed in the virtual world. A recent study by Shin et al. [35] has suggested that empathy and immersion must be measured together. The reasoning relates to the notion that empathy can only be fully experienced when a user is adequately immersed in VR. For VRLEs that include narratives or story telling or experiential application, empathy and embodiment are being included in measuring the effectiveness of VRLEs. 16.7% (n=4) of the studies measured embodiment [3, 34, 35, 37]. Phenomological proposal argues that empathy depends on embodiment [45]. Embodiment is the only feature tested by Ahn et al. [3] and Stiefs [37] while Shin et al.[35] and Shin [34] measured both embodiment and empathy along with other features. Ahn et al.[3] focused on measuring embodiment, experienced by users via spatial presence and body transfer, through questions with 5-point interval scales; Stiefs [37] tested embodiment using questionnaires; while Shin et al.[35] and Shin [34] used interviews, focus groups and questionnaires to measure empathy and embodiment. Shin [34] argues that empathy and embodiment should be measured, he stated that the cognitive processes by which users experience presence, and flow of the narrative will determine how they will empathize with and embody the story.

Felnhofer et al., [23] is the only study that used "elicit emotion" along with presence as a feature to evaluate their VRLE. Emotional experiences in turn are related to presence, another important concept in VR, which describes the user's sense of being in a VR environment. Specific application areas of VRLEs where eliciting emotion are useful are in experiential applications or narratives. In this study, 75% (n=3) of the samples measured presence but did not evaluate if they elicit emotion to its users.

3.2.5 Usability

Other studies on evaluating the effectiveness of VRLEs focus on usability and user satisfaction in general [13, 19, 35, 32, 38, 41]. Gil-Gómez et al. [25]. Priker et al[32] and Vergara et al.[41] created their own usability questionnaires while Butt et al. [14] used the System Usability Scale (SUS) which is a well established usability tool developed by Brooke [12]. Usability contributes to the overall user experience.

A closer look at the above-mentioned studies shows that the VRLE's features evaluated varies from study to study. In this case, the omission in measuring some of the VRLEs salient features to evaluate its effectiveness is confirmed given the limited number of features measured by the studies. It is not clear whether features evaluated can be reliably and systematically prescribed given the wide range of goals of VRLEs. However, it is possible to classify those features that have been measured in evaluating VRLEs to reveal common and distinctive features among them.

4. APPLICATION TO KOKODA VR

The salient features identified in this study were used in developing the instruments for evaluating the effectiveness of Kokoda VR. All features under the technical and human factors classification were used but only knowledge retention and mastery learning were used under the learning classification. Kokoda VR presents history and is a linear narrative application which does not assess students based on the task they perform rather on their understanding of the content.

Examples of questionnaires applying salient features in surveys or focus group:

Presence: *I am not aware of my real environment.*

Engagement: *I was excited to explore new things.*

Perception: *What are the best aspects of using Kokoda VR/ / Kokoda 360° Video?*

Elicit Emotion: *I felt sad when I saw the soldiers got sick.*

Empathy: *I understand why the Australian soldiers were feeling that way during the ceremony at the Kokoda village.*

Usability: *I find it easy to navigate the surroundings using the VR device.*

Immersion: *I felt stimulated by the virtual environment.*

Embodiment: *I felt in control when I picked up the objects in the virtual world.*

Mastery Learning or Retention Question: *Why were the men in the 39th battalion poorly prepared for war?*

In the development of the questionnaires, standards for developing questionnaires for educational research and handbook of questionnaire development were also used as a guide [4, 10, 43].

5. CONCLUSION

This paper attempts to identify salient features of an effective immersive non-collaborative VRLE. The review reveals interrelated features and their classifications that can be used as a guide in designing and evaluating this type of VRLE. The interrelatedness of these salient features discussed in the results show the importance of using them together to have a complete evaluation.

Presence is the most measured salient feature among the samples but is often measured with immersion. Presence is the psychological, perceptual and cognitive consequence of immersion. Perception's relationship with presence is that our perception is what we base our decisions on and mostly determines our sense of presence in the environment. Immersion, embodiment and empathy are also connected. Those studies that measured together these features argued that immersion has an impact on empathy and embodiment; and supported by studies that empathy depends on embodiment [45]. These two features are considered important especially for experiential and narrative story types of VRLE application.

This paper also showed that although the samples are evaluating VRLEs, not all samples measured learning. Effective VRLEs should ensure that there is some form of learning (understanding of the concept, knowledge retention, performance). Depending on what is the goal of the learning, VRLEs should measure a combination of the features identified. Motivation is often associated with any form of learning but in the samples, motivation is measured with performance. Engagement is a salient feature identified in this study that was associated with learning. As discussed in the results, users must be sufficiently engaged before any learning can occur but only few of the samples measured these features together.

Usability is also an important salient feature that must be measured to assess the quality of the student's experience using VRLE.

According to Whittemore et al [42], one implication of a review is to guide practice. The salient features identified in this review was the basis for the development of the instruments for evaluating the Kokoda VR case study. We anticipate that this instrument will provide us with detailed knowledge of the effectiveness of Kokoda VR, which will form the basis for designing immersive and non-collaborative VRLEs. The application of the classifications of the salient features can also be used in evaluating other types of VRLEs depending on focus areas of evaluation. For example, using a new VR technology using an existing VR application would most likely focus on the technology features while using the same immersive technology would look at focusing on learning or human factors' features. VR designers can also use these results to improve the learning effectiveness and human factors' features of their design.

6. ACKNOWLEDGMENTS

This research was supported in parts by funds received from the David A. Wilson Award for Excellence in Teaching and Learning, which was created by the Laureate International Universities network to support research focused on teaching and learning. For more information on the award or Laureate, please visit www.laureate.net

7. REFERENCES

[1] Abadia, R., Calvert, J. Tauseef, M.S. 2017. *Proposal: Analyzing the Effectiveness of Immersing Students in Interactive, Realistic Virtual Reality using Kokoda VR as Case Study*.. A Proposal to the David Wilson Award, Laureate Inc. Torrens University Australia.

[2] Addison, A. O'Hare, WT, Kerry, A., 2014. A Time and a Place to Learn: Can Learning Continue in Virtual Reality Long After Traditional Deliveries Have Exhausted Students Minds?.. *EDULEARN Proceedings(2014)* , pp 7618-7626

[3] Ahn, S. J. G., Bostick, J., Ogle, E., Nowak, K. L., McGillicuddy, K. T., & Bailenson, J. N. (2016). Experiencing Nature: Embodying Animals in Immersive Virtual Environments Increases Inclusion of Nature in Self and Involvement With Nature. *Journal of Computer-Mediated Communication*, *21*(6), 399–419. https://doi.org/10.1111/jcc4.12173

[4] Artino, A. R., La Rochelle, J. S., Dezee, K. J., & Gehlbach, H. (2014). Developing questionnaires for educational research: AMEE Guide No. 87. *Medical Teacher*, *36*(6), 463–474. https://doi.org/10.3109/0142159X.2014.889814

[5] Bamodu, O., & Ye, X. M. (2013). Virtual Reality and Virtual Reality System Components. *Advanced Materials Research*, *765–767*, 1169–1172. https://doi.org/10.4028/www.scientific.net/AMR.765-767.1169

[6] Bharathi, A. K. B. G., & Tucker, C. S. (2016). Investigating the Impact of Interactive Immersive Virtual Reality Environments in Enhancing Task Performance in Online Engineering Design Activities. In *Proceedings of the ASME 2015 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference* (Vol. 3).

[7] Bhargava, A., Bertrand, J. W., Gramopadhye, A. K., Madathil, K. C., & Babu, S. V. (2018). Evaluating multiple levels of an interaction fidelity continuum on performance and learning in near-field training simulations. *IEEE Transactions on Visualization and Computer Graphics*, *24*(4), 1418–1427. https://doi.org/10.1109/TVCG.2018.2794639

[8] Bloom,B. (1968). *Learning for Mastery*. Formative and Summative Evaluation of Student Learning. Mc-Graw-Hill.

[9] Borba, E. Z., Montes, A., De Deus Lopes, R., Zuffo, M. K., & Kopper, R. (2017). Itapeva 3D: Being Indiana Jones in virtual reality. *Proceedings - IEEE Virtual Reality*, 361–362. https://doi.org/10.1109/VR.2017.7892326

[10] Boyles, B. (2008). Using virtual reality and augmented reality in education, *1*(1), 2–4. Retrieved from https://www.usma.edu/cfe/Literature/Boyles_17.pdf

[11] Brancato, G., Macchia, S., Murgia, M., Signore, M., Simeoni, G., Blanke, G., Hoffmeyer-Zlotnik, J. H. . (2006). Handbook of Recommended Practices for Questionnarie Development and Testing in European Statistical Systems. *European Commission Grat Agreement*, 162. Retrieved from http://epp.eurostat.ec.europaipv6.eu/portal/page/portal/research_methodology/documents/Handbook_questionnaire_development_2006.pdf

[12] Brooke, J. (1996). SUS: A "quick and dirty" usability scale. *Usability Evaluation in Industry*. London: Taylor and Francis.

[13] Bruno, P., Ongaro, A., and Fraser, I., (2007). Long-term retention of material taught and examined in chiropractic

curricula: its relevance to education and clinical practice. *The Journal of Canadian Chiropractic Association*, 51 (1), 14-18.
[14] Butt, A. L., Kardong-Edgren, S., & Ellertson, A. (2018). Using Game-Based Virtual Reality with Haptics for Skill Acquisition. *Clinical Simulation in Nursing*, *16*, 25–32. https://doi.org/10.1016/j.ecns.2017.09.010
[15] Buttussi, F., & Chittaro, L. (2018). Effects of Different Types of Virtual Reality Display on Presence and Learning in a Safety Training Scenario. *IEEE Transactions on Visualization and Computer Graphics*, *24*(2), 1063–1076. https://doi.org/10.1109/TVCG.2017.2653117
[16] Buttussi, F., & Chittaro, L (2015). Assessing Knowledge Retention of an Immersive Serious Game vs. a Traditional Education Method in Aviation Safety. *IEEE Transactions on Visualization and Computer Graphics*, *21*(4), 529–538. https://doi.org/http://dx.doi.org/10.1109/TVCG.2015.2391853
[17] CAICT, Technology, C., & Co, H. T. (2017). Virtual Reality / Augmented Reality White Paper, (September).
[18] Carr, K. and England, R. (1995). *Simulated and Virtual Realities: Elements of Perception.*Taylor and Francis. ISBN-13: 978-0748401291
[19] Castro-Gonzales, D., Barbosa, LH. , Prada-Jimenez, V., Conde-Mendez, G. (2017). Design and development of an immersive virtual environment for teaching of the superposition of movements principle for engineering students. *Revista Educacion en Ingenieria*. Vol 12. Issue 23. Pp 101-108
[20] Chen, C. J. (2006). The design, development and evaluation of a virtual reality based learning environment. *Australasian Journal of Educational Technology*, *22*(1), 39–63. Retrieved from http://www.ascilite.org.au/ajet/ajet22/chen.html
[21] Chirico, A., Ferrise, F., Cordella, L., & Gaggioli, A. (2018). Designing awe in virtual reality: An experimental study. *Frontiers in Psychology*, *8*(JAN), 1–14. https://doi.org/10.3389/fpsyg.2017.02351
[22] Dang, B. K., Palicte, J. S., Valdez, A., & O'Leary-Kelley, C. (2018). Assessing Simulation, Virtual Reality, and Television Modalities in Clinical Training. *Clinical Simulation in Nursing*, *19*, 30–37. https://doi.org/10.1016/j.ecns.2018.03.001
[23] Felnhofer, A., Kothgassner, O. D., Schmidt, M., Heinzle, A. K., Beutl, L., Hlavacs, H., & Kryspin-Exner, I. (2015). Is virtual reality emotionally arousing? Investigating five emotion inducing virtual park scenarios. *International Journal of Human Computer Studies*, *82*, 48–56. https://doi.org/10.1016/j.ijhcs.2015.05.004
[24] Fominykh, M., Prasolova-Førland, E., Morozov, M., Smorkalov, A., & Molka-Danielsen, J. (2014). Increasing Immersiveness into a 3D Virtual World: Motion-tracking and Natural Navigation in vAcademia. *IERI Procedia*, *7*(2212), 35–41. https://doi.org/10.1016/j.ieri.2014.08.007
[25] Gil-Gómez, J. A., Manzano-Hernández, P., Albiol-Pérez, S., Aula-Valero, C., Gil-Gómez, H., & Lozano-Quilis, J. A. (2017). USEQ: A short questionnaire for satisfaction evaluation of virtual rehabilitation systems. *Sensors (Switzerland)*, *17*(7), 1–12. https://doi.org/10.3390/s17071589
[26] Goss, P., Sonneman, J., & Griffiths, K. (2017). *Engaging students: creating classrooms that improve learning. Grattan Institute Report No. 2017-01*.
[27] Hvass, J. S., Larsen, O., Vendelbo, K. B., Nilsson, N. C., Nordahl, R., & Serafin, S. (2017). The effect of geometric realism on presence in a virtual reality game. *2017 IEEE Virtual Reality (VR)*, 339–340. https://doi.org/10.1109/VR.2017.7892315
[28] Kim, M., Jeon, C., & Kim, J. (2017). A Study on Immersion and Presence of a Portable Hand Haptic System for Immersive Virtual Reality. *Sensors (Basel, Switzerland)*, *17*(5). https://doi.org/10.3390/s17051141
[29] Kuronen-Stewart, C., Ahmed, K., Aydin, A., Cynk, M., Miller, P., Challacombe, B., ... Popert, R. (2015). Holmium Laser Enucleation of the Prostate: Simulation-Based Training Curriculum and Validation. *Urology*, *86*(3), 639-646. DOI: 10.1016/j.urology.2015.06.008
[30] Lucas, J. (2018). Immersive VR in the construction classroom to increase student understanding of sequence, assembly, and space of wood frame construction. *Journal of Information Technology in Construction (ITcon)*, *23*(November 2017), 179–194. Retrieved from https://itcon.org/papers/2018_09-ITcon-Lucas.pdf
[31] Omrod, J.E. (2008). *Educational Psychology Developing Learners, p. 384-386.*
[32] Pirker, J., Lesjak, I., & Guetl, C. (2017). Maroon VR : A Room-scale Physics Laboratory Experience. *IEEE 17th International Conference on Advanced Learning Technologies*. https://doi.org/10.1109/ICALT.2017.92
[33] Rengganis, Y. A., Safrodin, M., & Sukaridhoto, S. (2018). Integration Head Mounted Display Device and Hand Motion Gesture Device for Virtual Reality Laboratory. *IOP Conference Series: Materials Science and Engineering*, *288*(1). https://doi.org/10.1088/1757-899X/288/1/012154
[34] Shin, D. (2018). Empathy and embodied experience in virtual environment: To what extent can virtual reality stimulate empathy and embodied experience? *Computers in Human Behavior*, *78*, 64–73. https://doi.org/10.1016/j.chb.2017.09.012
[35] Shin, D., & Biocca, F. (2017). Exploring immersive experience in journalism. *New Media and Society*. https://doi.org/10.1177/1461444817733133
[36] Slater, M., & Wilbur, S. (1997). A Framework for Immersive Virtual Environments (FIVE): Speculations on the Role of Presence in Virtual Environments. *Teleoperators and Virtual Environments*, *6*(6), 603–616.
[37] Stiefs, D. (2016). Embodied Experiment of Levitation in Microgravity in a Simulated Virtual Reality Environment for Science Learning. *IEEE Virtual Reality Workshop on K-12 Embodied Learning Through Virtual & Augmented Reality (Kelvar)*
[38] Tajiri, Keisuke; Setozaki, Norio. 2016. Development of Immersive Teaching Material using HMD and 3D Gesture Operation for Astronomy Education. *24th International Conference on Computers n Education (ICCE 2016): Think Global Act Local* Pages: 160-162
[39] Torraco, R. J. (2005). Writing Integrative Literature Reviews: Guidelines and Examples. *Human Resource Development Review*, *4*(3), 356–367. https://doi.org/10.1177/1534484305278283
[40] Usoh, M., Catena, E., Arman, S., & Slater, M. (2000). Using Presence Questionnaires in Reality. *Virtual Environment*, *9*, 497–503.
[41] Vergara, D., Rubio, M., & Lorenzo, M. (2017). On the Design of Virtual Reality Learning Environments in Engineering. *Multimodal Technologies and Interaction*, *1*(2), 11. https://doi.org/10.3390/mti1020011
[42] Whittemore, R., & Knafl, K. (2006). The integrative review : Updated methodology The integrative review : updated methodology. *Journal of Advanced Nursing*, *52*(5), 546–553.

https://doi.org/10.1111/j.1365-2648.2005.03621.x
[43] Wilson, K. L., Lizzio, A., & Ramsden, P. (1997). The development, validation and application of the Course Experience Questionnaire. *Studies in Higher Education*, *22*(1), 33–53. https://doi.org/10.1080/03075079712331381121
[44] Wonglorsaichon, B., Wongwanich, S., & Wiratchai, N. (2014). The Influence of Students School Engagement on Learning Achievement: A Structural Equation Modeling Analysis. *Procedia - Social and Behavioral Sciences*, *116*, 1748–1755. https://doi.org/10.1016/j.sbspro.2014.01.467
[45] Zahavi, D. 2011. Empathy and Directo Social Perception: A Phenomenological Proposal. Rev Philos Psychol 2(3):541-558

Proceedings of the SMART–2018, IEEE Conference ID: 44078
2018 International Conference on System Modeling & Advancement in Research Trends, 23rd–24th November, 2018
College of Computing Sciences & Information Technology, Teerthanker Mahaveer University, Moradabad, India

Security and Privacy in IoT based E-Business and Retail

Keshav Kaushik[1] and Susheela Dahiya[2]
[1,2]*School of Computer Science, University of Petroleum & Energy Studies, Dehradun, India*
E-mail: [1]officialkeshavkaushik@gmail.com, [2]susheela.iitr@gmail.com

***Abstract*—Interconnection of various "things" are resulting into many issues related to security and privacy of IoT platform. With the advancement in technology, many e-businesses and retail stores are using IoT based solutions for their sale, marketing, productivity and promotions. These IoT based solutions are much helpful in providing various benefits to owner and customers. However, these solutions are vulnerable to many security and privacy based concerns. This paper addressed the rise of cyber threats in IoT, Enterprise view of IoT for E-Business & Retail Security, advancements in e-business and retail because of IoT, flow of threat agents related to security and privacy concerns in e-business and retail.**

Keywords: Internet of Things (IoT), Security and Privacy, e-Business, Retail, IoT Architecture, Cyber-Attacks

I. Introduction

With the advent of Internet of Things (IoT), we can connect even those "things" to the internet that were actually not meant for Internet Connectivity. Due to inter-connectivity of these wide varieties of physical things, IoT domain is facing various issues related to security and privacy. Networking and sensing capabilities in any physical device are making them prone to latest cyber-attacks and privacy of users are at stake. More the number and variety of connected devices more will be the security and privacy problems related to them and more we will require the security patches and updates for those IoT devices.

E-Business and retail is growing by leaps and bounds due to growth of e-commerce websites and retail stores. We are installing more and more IoT devices in retail stores in order to provide users with a more realistic and friendly experience. However, these IoT devices involved in E-Business are becoming a soft target for hackers and unauthorized users. Imagine the level of chaos that could happen if those interconnected devices face a Ransomware attack or imagine a situation when the intelligent inventory management system of an e-commerce company is leaking out purchase details and order history of customers.

As we know that huge amount of data is collected continuously by IoT devices involved in e-business, which may include order history, purchase pattern of various customers, which may be very helpful in long run for analytics. Based on that, we can predict the trends and advancements in e-business. The data collected may also be used for research study of latest cyber-attacks and providing with some remedies for latest attacks. The paper will discuss about security assessment of IoT related solutions especially in E-Business which the need of hour. There are some critical security issues in IoT, which generally lead to the breach of privacy, and the same is highlighted in this paper. When it comes to E-Business security, there is a need of adapting a common standard architecture in IoT domain;

This paper is divided into five major sections. The first section gives a brief introduction about IoT, E-Business along with the need of privacy in IoT based E-Business. The rise of cyber threats in IoT are discussed in second section. The third section explains the architecture of IoT for E-Business & Retail Security. The fourth sections give a brief overview about the future of e-business and retail with IoT and fifth section gives overview of security and privacy issues in IoT related to e-business & retail. Lastly, it concludes with the findings from this research.

II. Rise of Cyber Threats in IoT

As many companies are increasingly digitizing their operations and adopting the internet of things platforms, software security experts warn of an increase in cyber-crime. Digital transformation is a significant focus for innovation and investment, triggering the creation of millions of new companies and is one of the drivers of the internet of things, which will see 50-billion devices connected to the internet by 2020, according to information technology giant Cisco [1].

A survey by Vodafone, which owns 65% of Vodacom [2], shows that 90% of respondents in SA believe that future success of any organization is critically based on Internet of Things, while 88% of them are of the opinion that seeing real success and value from the internet of things requires significant financial and time investment. More than 50 small to medium-sized enterprises in SA were interviewed for the survey [2]. Cyber-attacks are a numbers game. The attackers can be successful once, while defenders have to be successful multiple times. Companies that have been attacked often keep the breach under wraps to protect their reputations. In one of the most recent attacks, criminals in Japan who made 14,000 ATM withdrawals stole R350m from Standard Bank.

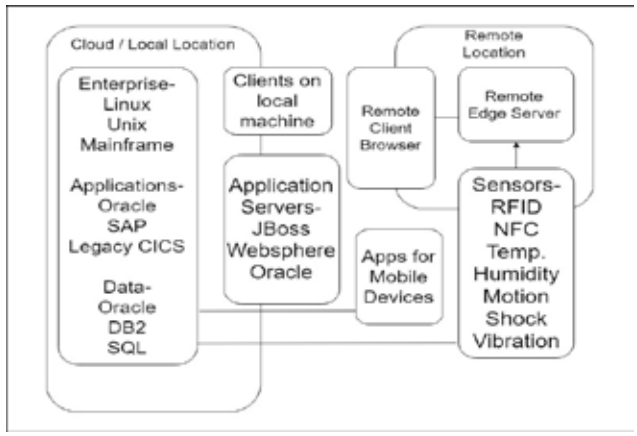
In the recent past, expensive resources like bandwidth, memory, address space, servers were targeted by massive botnets in their costly hosting environment. Internet of Things revolutionizes these botnets [7]. Today, attackers can crack the IoT device default credentials by spending small amount of work and money. So, attackers tar-get these devices instead of building their own botnets in costly hosting environment. According to Symantec, the principal cyber threats to Internet of Things includes Denial of Service, Botnets and Malware based attacks, data breaches, inadvertent breaches, weakening perimeters. In 2014, OWASP [5] also released its top 10 vulnerabilities for Internet of Things, which proves the considerable indulgence of cyber security attacks in IoT, and they need to be tackled in a proper way.

III. Enterprise View of IoT for E-Business & Retail Security

With the drastic growth in e-business domain, the demand for technical advancements is also increasing and IoT supports it. Important departments of e-business like payment, logistics, inventory etc are affected by applications of IoT [9]. IoT technologies like ZigBee, RFID, NFC, WiFi-Direct, low-energy radio protocols are applied in all the departments of e-business. With the application of these advance technologies, there comes the security concern e.g Payment made in e-business through IoT technologies. It is also proven that for the growth, cut throat competition and promotion of e-business we need to take help of such IoT technologies. The efficiency of e-business ecosystem will be improved by IoT technologies. As it is improving the customer's experience, reducing the overall cost of operations, giving a better look, feel and brand value. IoT is providing the retailers in e-business with a more realistic, flexible, efficient and holistic experience as they can personalize many things and engage customers for better relation and more benefits. Based on shopping habit and pattern of customers, companies can propose some attractive offers, which may ultimately improve the sale of items.

There are various security and privacy concerns as well with IoT application in e-business. For example, as we have mentioned that some offers or schemes are re-leased by monitoring the shopping pattern or shopping habit of various persons. Now, this monitoring itself raises various privacy concerns as this data may be leaked over the internet. It may expose the details about various products that a particular person is using which will be helpful for attacker in spear phishing attack. Some secure con-strained devices used in e-business are vulnerable to side channel attacks, such as power analysis attack, which is actually used to reverse engineer the applied algorithm. Sometimes authentication and authorization itself can a potential vulnerability in IoT based devices. Default passwords or weak passwords are open invitation for attackers to exploit IoT devices. When talking about installing the security updates in IoT devices they need to be properly taken care of. In e-business, we cannot afford a considerable downtime for a IoT solution as it me affect the business directly. Personally Identifiable Information can be decoupled from IoT data payloads if proper data privacy is not implemented in IoT devices. In addition, data integrity can also be compromised if proper checksums and digital signatures are not properly implemented.

Fig. 1 shows the IoT Enterprise architecture. In this architecture, the enterprise may be mainframe, Linux that runs on various applications based on SAP, Oracle etc. The backend is supported by databases like SQL, Oracle, and DB2. Clients on local ma-chine can interact with the remote machine present at remote location with the help of browser at the client side. They also use various application servers like Web-Sphere, JBoss, Oracle, and the same can be used with the help of applications present on mobile devices. Various sensors like humidity, motion, shock, vibration are used for communication. They use latest technologies like RFID, NFC, ZigBee, LowPAN, CoAP for smooth interaction among various things [8].



Fig, 1: IoT Enterprise Architecture in e-Business and Retail

IV. Future of e-Business and Retail with IoT

IoT is affecting the e-business and retail in a very significant way. IoT gives ability to retailers to connect people with things. It also giver retailers an insight on who their customers are, how they are involved in product purchase, how their products are performing and how retailers can engage their customers with existing and new products. In addition, IoT helps retailers to introduce new products with more functionalities depending upon the current needs. The main aim of introducing IoT in retail and e-business is to create a rich and more realistic shopping experience. Techniques like Augmented Reality (AR) [10], Virtual Reality (VR) and wearable technologies are helpful in customer engagement. Various analytics tools are helping the retailers to collate and collect more data in a real time environment. Smart beacons installed at the retail stores send notifications to customers automatically when they

enter the store. Advanced devices like Smart mirrors allows customers to try the clothes virtual-ly and thus increasing the shopping experience without forcing them to travel to the changing room. In monitoring of inventory at retail stores Smart Shelves are used which notifies the manager if the inventory is running low.

V. Security and Privacy Issues in IoT Related to e-Business & Retail

As more and more devices are being connected, the chances of security and privacy breaches are being increased. One more reason for such kind of breaches is ignorance or laziness or users, they are not installing the security updates frequently. IoT devices can generate sheer amount of data. According to a report by Federal Trade Commis-sion with title "Internet of Things: Privacy & Security in a Connected World", around 10,000 households can generate 150 million discrete data points every day. There are some companies, which can use the data collected from customers in making some important decisions. For example, an insurance company may collect some infor-mation about your driving habits through a IoT enabled car when calculating your insurance rate. This is possible because of ignorant nature of customers who generally do not read End User License Agreement (EULA). Eavesdropping may also raise some privacy concerns. For example, a connected device may intercept the nearby-unencrypted traffic and can get television shows someone is watching at that mo-ment. There are various kinds of privacy risks involved, some of these risks include – collection of sensitive information, like accurate geolocation, financial account num-ber, issue related to healthcare like physical conditions over time, health habits. In fact, an existing smartphone sensor can be used to predict a user's mood, personality type, smoking habits[4], stress level, types of physical activities, sleep pattern. Such kind of critical information can be used in an unauthorized way if not properly han-dled. Some companies may also use such type of information for making crucial employment decisions, for insurance and for selling credit cards. Some insurance companies also use driving habits of a person in order to set the insurance rate.

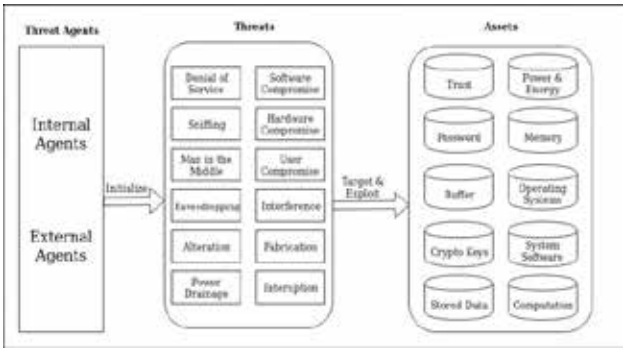


Fig. 2: Security and Privacy Flow for IoT Related to e-Business and Retail

Talking about security and privacy flow in IoT (Fig. 2.) [3], there can be internal and external agents, who actually are responsible. These agents can harm personal privacy, cyber thefts, financial transactions, data breach etc. through various kinds of cyber-attacks. These attacks can target- memory, operating system, system software, power & energy of an IoT device etc. Compromise of any type on IoT services, applications and devices can serve as an entry point for data theft and unauthorized access. These cyber-crimes can be on individuals as well as on organizations. These attacks are more intrusive, difficult to control and are sophisticated nowadays which are creating a lot of challenges for cyber security experts.

In general, these security and privacy threats in IoT are categorized into four domains [3]:

1. Application Based Threats
2. Cloud Based Threats
3. Hardware devices based Threats
4. Communication based Threats

A. Application Based Threats

In this domain, threats are targeted on various IoT based applications. These attacks arise due to insecure applications, mobile or desktop based. Some applications are not password protected which results in severe privacy and security concerns. Poor or low security configuration of various applications invites security concerns. Multiple users accessing an application may also lead to security issues if roles of those users are not well defined. Such kind of vulnerabilities may lead to collection of personal information and its dissemination over insecure public network.

B. Cloud Based Threats

Various IoT applications and services are using cloud for communication and data sharing. If cloud is poorly configured over SSL/TLS, then it may lead to severe cyber-attack. Sometimes clouds are also targeted for SQL Injection attacks because they do not use proper sanitization and input validation. Out of date legacy devices invites attackers for cloud-based attack. Insecure cloud interface and in-cloud data leaks also add up to various attacks which fall under this category.

C. Hardware Devices Based Threats

Sometimes IoT hardware devices are targeted by attackers because of insecure open external ports. Malicious software updates and outdated legacy devices are also vulnerable for any IoT devices. Use of default passwords for a long time is also a potential threat to security and privacy for IoT devices.

D. Communication Based Threats

Improper use of IoT protocols may lead to communication based IoT threats. Distributed Denial of Services (DDOS) is a potential threat, which may affect

the security of any organization. Use of poor encryption algorithms, vulnerable services over insecure network, open ports between peer to peer networks are potential threats which falls under this category.

Based on above threats, there are various recommendations, which will help in securing IoT ecosystem. These recommendations are tabulated below in Table 1:

TABLE 1: SECURING IOT ECOSYSTEM IN E-BUSINESS AND RETAIL

Sr. No.	Threats	Recommendations for Securing IoT Ecosystem
1.	***Application Based Threats***	• Implementing secure coding and testing techniques while developing IoT applications. • Use of patterns / Pins / Fingerprints while developing applications so that they lock out apps and prevent them from unauthorized access. • Use of Dynamic Application Security Testing and Zed Attack Proxy will prevent apps from attacks. • Monitoring of access and permissions to IoT applications based on role of user.
2.	***Cloud Based Threats***	• Use protocol filtering at cloud. • Implement data protection at in-cloud level. • Proper access controls for configuration and device identity using Trusted Platform Module. • Protect cloud infrastructure with SSL/TLS.
3.	***Hardware devices based Threats***	• Implement strong security at sensors level. • Updating and securing outdated legacy devices. • Encryption on hardware updates and patches. • Use advance security on gateways and firewalls. • Access privileges and control must be role based. • Device booting must be secured
4.	***Communication based Threats***	• Use of secure services for communication. • Rate Based Intrusion Prevention for monitoring of security logs. • Use of advance IDS/IPS. • FIPS, SSL, DTLS, SSH, latest ISO standards for encrypting data at rest and data in motion.

VI. CONCLUSION

Internet of Things offers a variety of applications in various domains. However, when talking about e-business and retail, IoT completely changes this industry very significantly in the past years. This change is phenomenal, it arises various open is-sues related to security and privacy, which needs to be addressed on priority by re-search community in order to make e-business and retail secure in reference with IoT. This paper has highlighted the IoT enterprise architecture in e-business and retail. In addition, major security and privacy threats of IoT in e-business and retail with their recommended solutions are addressed. We also discussed the flow of security and privacy threat agents pertaining to IoT in e-business and retail. In future, research on the security and privacy in IoT based e-business and retail will remain a hot issue.

REFERENCES

[1] Evans, D., “The internet of things: How the next evolution of the internet is changing everything,” CISCO white paper, 1(2011), pp.1-11, 2011.

[2] Cybercrime ‘will rise’ with internet of things, https://www.businesslive.co.za/bd/life/gadgets-and-gear/2016-11-22-cybercrime-will-rise-with-internet-of-things/ (Accessed on: 06/10/2018)

[3] IOT CONNECTED WORLD : SECURITY AND PRIVACY https://www.infosys.com/industries/communication-services/white-papers/Documents/IoT-connected-world.pdf (Accessed on: 06/10/2018)

[4] Federal Trade Commission. “Internet of Things: Privacy & security in a connected world.” Washington, DC: Federal Trade Commission (2015).

[5] Internet of Things: How Much are We Exposed to Cyber Threats? https://resources.infosecinstitute.com/internet-things-much-exposed-cyber-threats/(Accessed on : 06/10/2018)

[6] L. Atzori, A. Iera, and G. Morabito, “The internet of things: A survey,” Computer networks, vol. 54, no. 15, pp. 2787–2805, 2010.

[7] The Hunt for IoT: The Rise of Thingbots, https://www.f5.com/labs/articles/threat-intelligence/the-hunt-for-iot-the-rise-of-thingbots

[8] Singh S and Singh N.,”Internet of Things (IoT): Security challenges, business opportunities & reference architecture for E-commerce”. InGreen Computing and Internet of Things (ICGCIoT), 2015 International Conference on 2015 Oct 8 (pp. 1577-1581). IEEE.

[9] Xu, Xiaoming. “IOT Technology Research in E-commerce.” Information Technology Journal 13.16, pp: 2552-2559, 2014

[10] The future of retail through the internet of things: https://www.intel.com/content/dam/www/public/us/en/documents/white-papers/future-retail-through-iot-paper.pdf (Accessed on : 06/10/2018)

Security techniques using Enhancement of AES Encryption

[1] Meetender , [2]Nirbhay Kashyap, [3]Archit Aggarwal, [4]Tanupriya Choudhury
[1],[2,3]Amity University Uttar Pradesh, [4]University of Petroleum & Energy Studies (UPES),Dept. of Informatics,School of Computer Science,Dehradun
[1]mitenderadhana3@gmail.com, [2]nkashyap@amity.edu,[3]archit.aggarwal1508@gmail.com, [4]tanupriya1986@gmail.com

Abstract In current time there is a need of a network system which can be used in diverse application. Whenever there is new security threat arrives in system it's initially very difficult to detect. Encryption algorithms are very popular for network communication of data in information security system. There are different methods of protection of data which is shared on channel. S-box is very popular design to enhance the strength of cryptographic data. . The retardation is the flaw of the current one dimensional design in the S-box.

Keywords*:* AES, AES Enhancement, Advanced Encryption Standard, Encryption, Security

1. INTRODUCTION

The rapidly developing assortment of remote correspondence clients has prompted expanding interest for safety efforts and gadgets to protect client data transmitted over remote channels[1]. Two sorts of crypto legitimate frameworks have been produced for that reason symmetric and awry cryptosystems. Symmetric cryptography like inside the Data Encryption Standard, DES and Advanced Encryption Standard (AES), utilize a similar key for the sender and collector, each to encode the first message and unscramble message figure content. Deviated cryptography, as inside the Rivest-Shamir-Adleman (RSA) utilizes totally unique keys for encoding and deciphering, disposing of the key trade disadvantage. Symmetric cryptography is more suitable for the encoding of an outsized measure of data. The AES calculation sketched out by the National Institute of Standards and Technology in US has been broadly acknowledged to switch DES in light of the fact that the new symmetric encryption calculation. The AES calculation could be a symmetric square figure that procedures information pieces of 128 bits utilizing a figure key of length 128, 192, or 256 bits. Each datum square contains a 4×4 exhibit of bytes called as the state, on that the AES calculation is connected[2]. The proposed calculation contrasts from standard AES in light of the fact that it has 200 bits piece size and key size each. Number of rounds is steady and up to 10 in this calculation. The age of key development and substitution enclose are done a comparative technique as in standard AES piece figure. AES has 10, 12 and 14 rounds for 128-bit keys, 192-bit keys and 256-bit keys separately.

2. PROPOSED WORK

2.1 AES Algorithm

AES algorithm uses 10,12,14 rounds of function depends upon the size of the key[3]. Each rounds consists of four functions, if aes contains n rounds then n-1 rounds contains all the 4 operations but n rounds uses only 3 rounds in encryption as well as decryption[4]. In Encryption step n round does not contain mix columns rounds. The block diagram of AES is as shown below

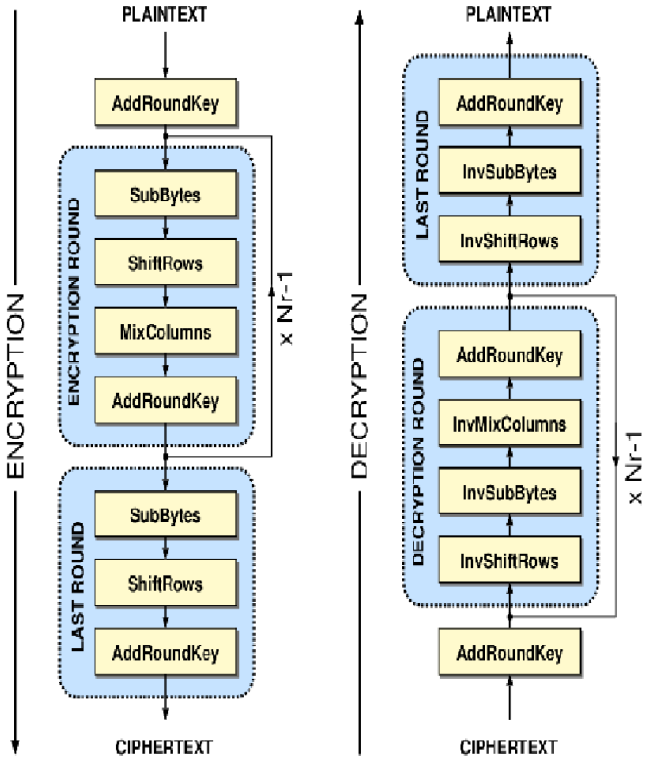


Fig 1. Block Diagram of AES

Keys	No. of rounds
128	**10**
192	**12**
256	**14**

Table1. Relationship between keys and no. of rounds

Each Round Of Encryption Consists of 4 Operations namely Sub-Bytes, Shift Rows, Mix Column and Add Round Key. Each Round of Decryption Consists of 4 Operations Which are:

1) Inverse Shift Rows
2) Inverse Sub Bytes
3) Add Round Key
4) Inverse Mix Columns.

1) Sub Bytes

The first step is Sub Bytes, In this we substitute the value of matrix with Substitution Box, there is entry in Substitution box for each byte. The Inverse operation is known as inverse Sub Bytes which is used in Decryption step [5].

2) Shift Rows

Second function is Shift Rows, as name suggests we shift the rows in matrix toward left in circular manner. The first row will be remaining same in new matrix. The second Row will be shift by 1, Third row will be shifted by 2 and last row will be shifted by 3. The reverse operation is known as inverse Shift Rows

3) MixColumn

The 3rd operation is Mixcolumns, as name suggests in this step we mix the column.

We multiply the given state matrix with fix matrix and the output is then act as input to the next operation.

The inverse operation of mixcolumn is inverse mixcolumn operation and there is different matrix which is used in reverse operation [6].

4) AddRoundKey

The Final operation of single round of AES is called AddRoundKey. In initial 3 operations we did not use any keys so that operations can be reversible without knowledge of keys, so security is very less in that 3 steps, In AddRoundKey step we do the Xor operations of state table with the matrix of 4*4 of keys. if there are n rounds in AES operation then there will be n+1 AddRoundKey. We do AddRoundKey before any round which provides more security. The Key matrix is given by the key Distribution system.

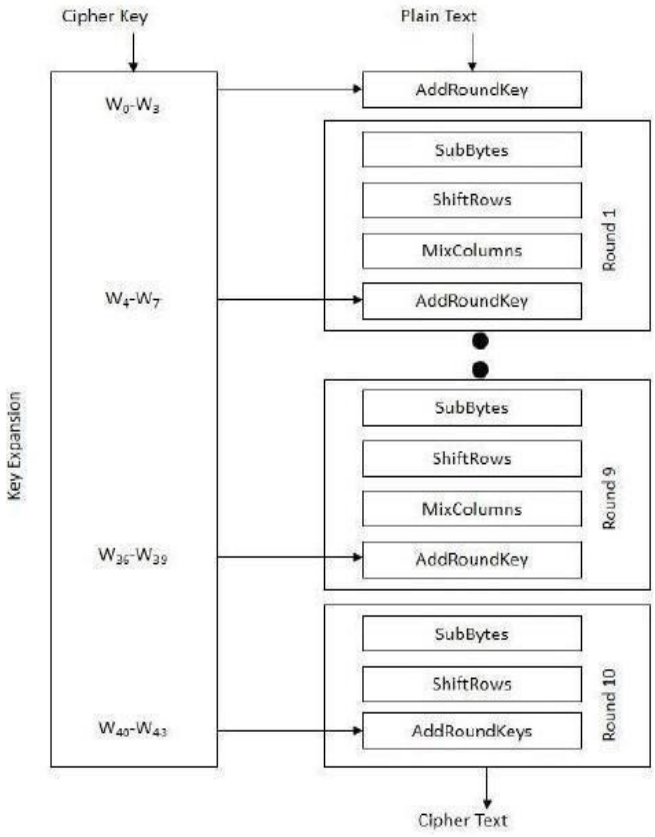


Fig 2. AES Diagram

Data Security with Tokenization

Data security model using tokenization can be developed which replace important data, in any financial database, with a token. This token is encrypted and stored in central data warehouse[7]. Proper unlocking mechanism is designed to be used at receiver end. Designed token is now ready to be passed in the network, leaving the encrypted data the token represents securely stored in the data vault.
Token can be produced in two ways –

i) The first information can be created utilizing token
ii) The first information can't be created utilizing token.

Tokenization can upgrade the assurance of delicate information by offering a token-based information.
At the end of the day, rather than keeping up ciphertext and a related key (ID) inside the Organization's information storage, a solitary token is put away and utilized as a pointer to the scramble an incentive in the vault. In managing an account framework, a charge card number, for instance, is supplanted inside the shipper's stockpiling condition by a token which is created such that it can't be

connected back to the first information component. A safe cross-reference table is built up to permit approved query of the first esteem, utilizing the token as the record. Encryption instruments and secure key administration supplements this approach by ensuring the first incentive inside this condition. To any individual who doesn't have approval to get to the vault, the token esteem is absolutely aimless; it's quite recently arbitrary characters. Proposed Work Plan in Tokenization depicted below.

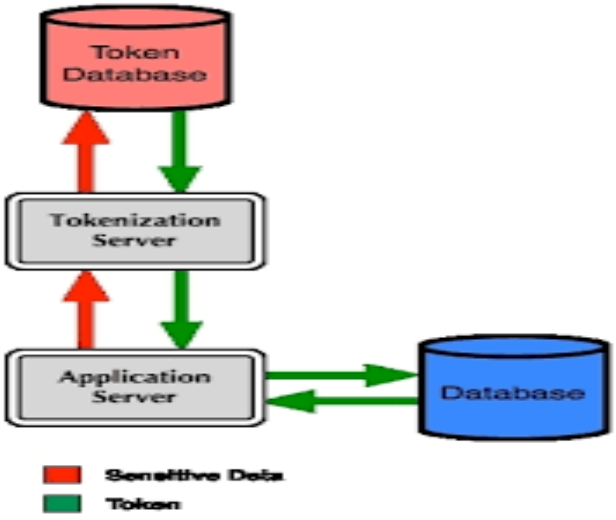


Fig 3: Proposed Work Plan in Tokenization

1. **Identification of Sensitive Data:** ID of delicate information is the sole obligation of the cloud shopper. There are diverse sorts of information regarding secrecy, protection or affectability. Ordinarily every one of the information of an association is not 100% delicate to protection and classification. Before outsourcing information by cloud purchaser, it is conceivable to distinguish level of classification of their information. In some Government Organizations, SSN- Standardized savings Number of the worker is exceptionally classified as exposure of such characteristic is not fancy by the representative or the Government don't prescribed to make such trait open. Another quality in particular, Telephone is to some degree delicate as the majority of the workers won't not yearning to reach number open. Nonetheless, the rest of the qualities, for example, Worker Code, Name, Assignment, Address and so forth are not all that touchy contrast with other attributes. Even though representative's name is exceptionally private, this is up to the manager of the association to characterize the sorts of information as far as classification, affectability or security.

2. **Segregation of Highly Confidential Data**: In the wake of breaking down information of the association, the information chairman can characterize levels of privacy which can be connected to the association's information. At that point, the distinguished information can be mapped with the characterized level of confidentialities Different security instruments, for example, encryption, tokenization and so forth could be chosen for each level of information. At that point, information with very secret will be isolated with certain connecting component for information tokenization. Assume, the System Administrator characterizes three levels of classification and same is connected to the distinguished information of table at Fig. 1, the accompanying table demonstrates the move to be made up for information assurance. At that point, the accompanying table containing "SSN" will be isolated.

Confidentiality protection

level	field	action
01	SSN	Tokenization
02	Phone	Encryption
03	Name,	No protection
04	Employee Code,	Encryption

Table2. Level of Fields, Action to be taken for Information

3. SECURITY CARE OF SENSITIVE DATA:

Once the information recognizable proof and isolation of exceedingly secret information is finished, the following obligation is safety efforts to be connected on the information of the association. At the same time, the information executive can choose one of the accessible calculations to encode the information whichever is important and tokenize the exceedingly touchy information. The tokenize information i.e. the surrogate esteems will supplant the first information. In first CSP condition, just the surrogate estimations of exceedingly secret information are accessible alongside related data. On the off chance that some person takes the information or break the information, the first information won't touch, just the surrogate, stunning esteems will be influenced. In second CSP condition, there is no genuine significance of exceptionally private information as the information is not related with other applicable data. Now we will use encoding on the selected data to compress it so we will use Huffman Encoding here. Huffman Encoding. Information weight acknowledges a fundamental part in PC structures. To transmit information to its goal speedier, it is fundamental to either build the information rate of the transmission media or basically send less information. The information weight is utilized as a bit of PC structures. To make the PC sorts out snappier, we have two alternatives i.e. one is to by one means or another improvement the information rate of transmission or some way or another send the less information. In any case, it doesn't gather that less data ought to be sent or transmitted. Data must be done at any cost.

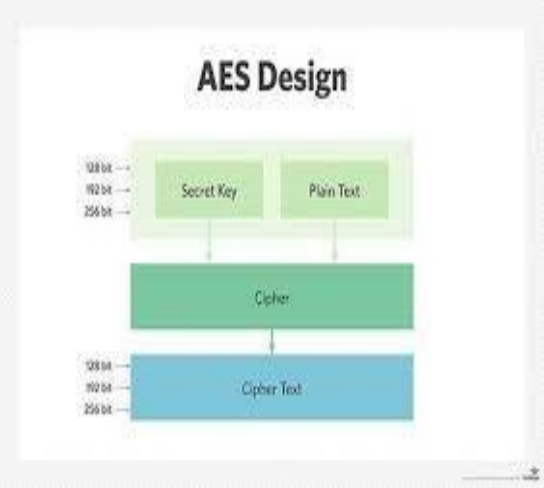


Fig 4: AES design

The Output of the Huffman Encoding Will be in compressed form. Now it will work as a input to the Aes encryption process here. The further Steps are mention below

1. Take the original s- box and find its 1`s complement
2. Then perform the xor operation between original s-box and 1`s compliments of s-box
3. The output will be come as virtual s-box which will be used here for encryption process
4. The reverse process will takes place for decryption process

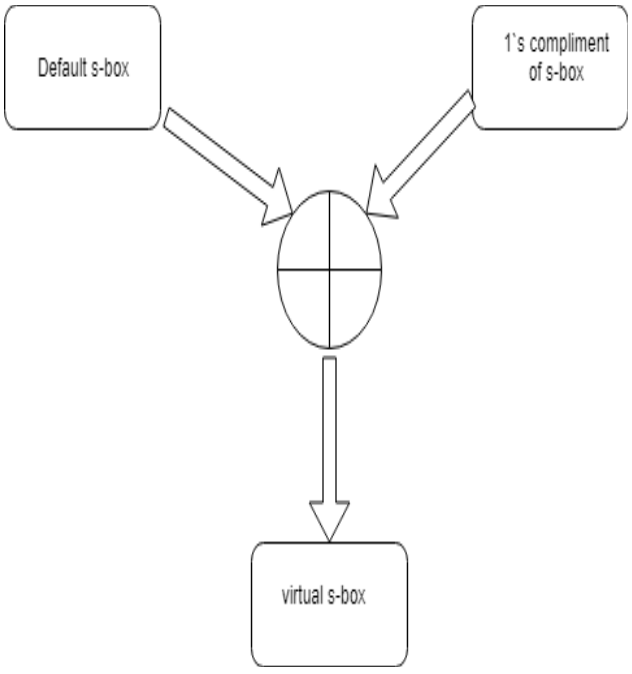


Fig 5. Virtual s-box implementation

5. At the end we can also use the biometric Security For Encryption and decryption process.

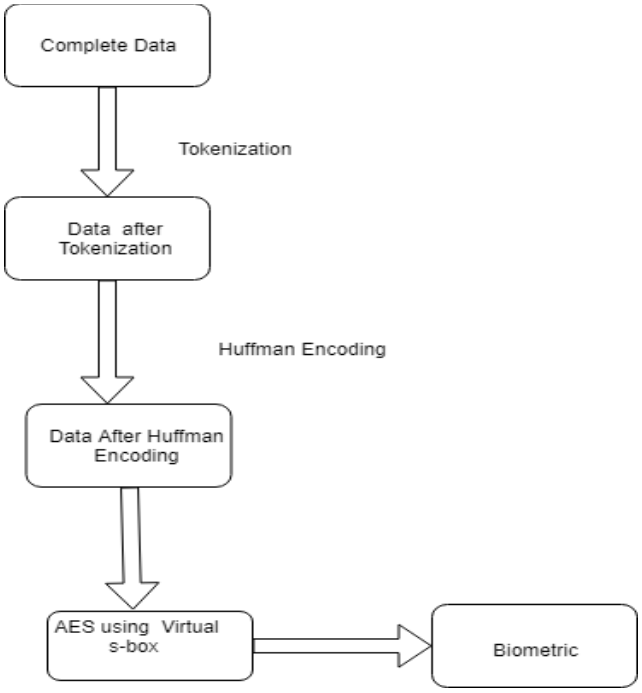


Fig 6. Flow diagram of proposed method

4. CONCLUSION

In this revolutionary age of information, security of that information is and should be the primary concern for any individual. The threat to the user's data remains constant and thus is of crucial importance for better algorithms to be developed.

The algorithm developed presents a large scale application base in all domains of data. The type of data to be encrypted is of negligible importance in regard to the algorithm. The algorithm successfully presents with the expected behaviour and results.

Comparison is done after doing theoretical computation and experimental analysis. The projected AES computation with crossover method will be a useful method in development of soft security in communication transmission. We have developed a better method for maintaining quality in AES to build transmission in Cipher text. Based on sample data taken and experimental result the AES has been designed. Thus we can successfully conclude that a better algorithm has been developed using the approach presented through this paper which is capable of providing a better level of encryption and overall security to sensitive information.

REFERENCES

[1] Jumaah B , Alahmad M.A. and Alshaikhli ,"Protection of the Digital Holy Quran Hash Digest by Using Cryptography Algorithms," Advanced Computer Science Applications and Technologies (ACSAT), 2013 ce on ,Vol. 244 , pp 249, 23-24

[2] Xianping Wu, Huy Hoang Ngo, Campbell Wilson and Phu Dung Le , “Dynamic Key Cryptography and Applications”, Faculty of Information Technology, Monash University , Australia.

[3] Zinan Chang, Fei Shao and Yi Zhang. 2010. "AES Encryption Algorithm Based on the High Performance Computing of GPU," International Conference on Communication Software and Networks, 2010, vol., no., pp. 588,590, 26-28

[3] W. Sun. and Feng, W. Lu, “Secure Binary Image Steganography Based on Minimizing the Distortion on the Texture”, IEEE transactions on Information Forensics and Security, 2015

[4] D. A. Tarah and A. Janadi, “AES immunity Enhancement against algebraic attacks by using dynamic S-Boxes,” 3rd International Conference on Information and Communication Technologies: From Theory to Applications, pp. 1-6.

[5] B. Li, X. Li, B. Yang and T. Zeng. “General framework to histogram shifting-based reversible data hiding”, IEEE Transaction on Image Processing”, 2013.

.

[6] R. Shakerian, S.H.Kamali, M. Rahmani ,M. Hedayati, “A new modified version of Advanced Encryption Standard based algorithm for image encryption”, IEEE International Conference on Electronics and Information Engineering (ICEIE), 2010.

[7] Subhashis Maitra., Subijit Mondal, “Data security - modified AES algorithm and its applications”, ACM SIGARCH 2014

Sentiment Analysis of User Entered Text

[1]Sourav Mehra,[2]Tanupriya Choudhury

[1,2]University of Petroleum & Energy Studies (UPES),School of Computer Science,Dehradun

[1]Souravmehra2018@gmail.com,[2]tanupriya1986@gmail.com

Abstract: **In this paper, classification algorithms (SVM and Naïve Bayes) are compared based on their classification accuracies on different datasets. They are then used to classify texts as positive or negative. For training and testing the algorithms, we will use the IMDb Large Movie Review Dataset, which contains 25,000 polar movie reviews each for training and testing and a movie reviews dataset provided by Cornell University containing 1,600 polar reviews each for training and testing.**

Keywords**:** ***Machine Learning, Sentiment Analysis, Support Vector Machine, Naïve Bayes***.

1. INTRODUCTION

Sentiment Analysis means analyzing the sentiments behind a text, tweet, etc. using Natural Language Processing (NLP) techniques and determining whether they are positive, neutral or negative. Positive emotions such as cheerfulness, encouragement, etc. are clubbed together while emotions like jealousy, sadness, aggressiveness, etc. come under negative emotions. Sentiment Analysis can also be used in political and commercial fields to understand the sentiments and actions of the public. The results of this analysis can then be used to advertise to specific people, like those who are supportive of some parties, or those people who are looking to buy specific products.

It falls under the domain of **Pattern Classification**. Pattern Classification means discovering patterns in a large dataset, either automatically (using unsupervised algorithms) or semi-automatically (using supervised algorithms) and classifying them into different classes. It relies on techniques of **NLP** to extract the important features from the data and on **ML** techniques for accurately classifying the input (user entered sentence) as positive or negative.[1]

Most words and phrases of most languages naturally tend to have a positive or negative undertone. Words like ‘excellent’, ‘outstanding’, ‘extraordinary’, etc. tend to be associated with positive emotions while words like ‘evil’, ‘disgusting’, ‘ugly’, etc. are categorized as negative ones. So if a sentence contains a greater number of positive words than negative ones, it will be categorized as a positive sentence; otherwise it will be categorized as a negative one. This approach is called sentiment polarity.[2]

2. LITERATURE REVIEW

2.1 Supervised Learning

A majority of practical machine learning projects use supervised machine learning. In supervised machine learning, the system tries to learn from existing examples, which we call the dataset.

Supervised learning is a type of learning where the system learns from the dataset in which we have both, the input data and the corresponding output data. The system studies the inputs and their corresponding outputs, learns the patterns behind them, and once when it is trained, uses what it has learnt to provide highly accurate results. This is similar to how children learn; a teacher (supervisor) teaches them with data like numbers, information, formulae, etc. and how to come up with solutions when a problem is presented. After learning from this, the children find solutions to similar problems by themselves.[7][8][9]

Supervised learning basically are of two types, namely classification, and regression.

- **Regression**: Regression problems are those problems where the output is a real value, such as ‘Rupees’, ‘weight’ or ‘height’.
- **Classification**: Classification problems are those problems where the output is a category or a group, such as ‘open and ‘close’ or ‘positive’ and ‘negative’.

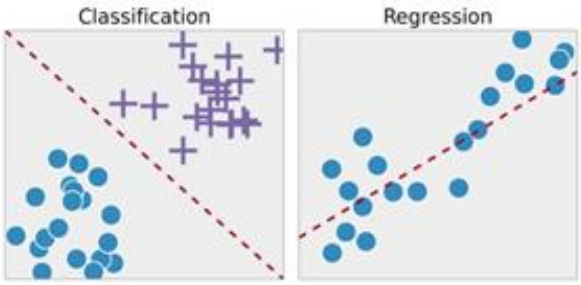


Figure 1: Classification vs Regression

2.2 Unsupervised Learning

Unsupervised learning refers to the learning where the user has to study the data and learn by themselves. There are no outputs available to the user. They have to find the pattern in the data by themselves, and then use that pattern to provide results.

This is similar to how researches are conducted. Scientists and researchers conduct experiments and collect and record readings generated in the experiment. They then study the patterns behind the collected data, to come up with new formulas and findings.

Unsupervised Learning is called so because there is no teacher to teach the system and there are no correct answers. Algorithms discover and interpret the structure in the data on their own.

Unsupervised learning basically are of two types, namely clustering and association.

- **Clustering:** A clustering problem is where we find groups in data, such as groups of customers according to their purchasing behavior.

- **Association:** A learning by association problem is where you discover patterns in data that describe large portions of it, such as people who are rich tend to have more weight.[3]

Figure 2: Clustering vs Association

We will be focusing on Supervised Learning in this paper.

3. METHODOLOGY

The process of creating a sentiment classifying application can be divided into the following steps/modules:

i Collecting Data
ii Manual Labelling
iii Data Cleansing
iv Classification
v Application

3.1 Collecting Data

Data was collected from two sources, namely from IMDb *Large Movie Review Dataset,* which contains 25,000 polar movie reviews each for training and testing and a movie reviews dataset provided by Cornell University containing 1,600 polar reviews each for training and testing. We then make three datasets out of them of sizes 1,600, 10,000 and 25,000 reviews. Both of the original datasets are already segregated into positive and negative categories. One can also create their own dataset by scraping reviews from movie review sites. The data collected in this step is later used to train and test the model.

3.2 Manual Labelling

The data which we are using has already been labelled. If someone wants to create their own dataset, they need to manually segregate the dataset into positive and

negative classes so that we can train the model on the data. This is how the datasets which are being used have been prepared, by scraping the internet for reviews, followed by manual segregation of those reviews into positive and negative classes.

3.3 Data Cleansing

The data in the dataset is just the raw data collected directly from review sites without any data processing, meaning that the data still contains punctuation marks, unnecessary whitespaces, numbers and special characters among many other entities that are not required for modelling. In this step we clean the data of all the entities mentioned above, so that we can fit the raw data into the model.

Given below are some statistics for the datasets used before and after data cleansing (preprocessing).

	Dataset		
	1,600 reviews	**10,000 reviews**	**25,000 reviews**
Total number of tokens	1,344,428	2,340,583	5,844,680
Number of stop words	530,117	1,092,611	2,729,715
Number of punctuation marks	225,220	532,182	1,327,499

Table 1: Data before cleansing

	Dataset		
	1,600 reviews	**10,000 reviews**	**25,000 reviews**
Number of unique tokens	44,090	81,608	138,430
Number of stop words	0	243	261
Number of punctuation marks	0	0	0

Table 2: Data after cleansing

3.4 Classification

It is a technique for segregating data (or any other collection of entities) into different categories according to their inherent properties or some preexisting patterns behind them. The aim of the project is to train a model that classifies the user input (text) into two sentiment classes (positive and negative) accurately.

Consider the sentence “My car is far better than his hotel”. **General Sentiment Analysis (GSA**) (which we will be using in this project) will categorize the above sentence as negative. **GSA** analyses the sentiment of the entire text as a whole. Therefore for the above example, since there is an overall negative undertone, a good GSA classifier would identify it as negative. [4]

To make classification easier, every positive and negative word of the training dataset are mapped to ‘1’ and ‘0’ respectively. SVM and Naïve Bayes classifiers will be used for classifying the reviews as either positive or negative and then we’ll be comparing the classifiers based on their accuracies.

3.5 Application

All of the above steps are compiled into an application/program which takes users’ inputs in the form of texts and classifies them as positive or negative. Initially, the user enters the text, then the text gets cleaned of punctuation marks, white spaces, emoticons, etc. Then the remaining words are determined to be positive or negative. If more positive words than negative ones are present, the text is classified as positive, otherwise it’s classified as negative.

4. SVM vs NAÏVE BAYES

4.1 SVM

SVM (Support Vector Machine) is a supervised machine learning algorithm that can be used in classification as well as regression problems. It is a hyperplane based classifier, i.e., it uses hyperplanes to perform classification. It gives an optimal hyperplane as output when labeled training data is given as input. It then uses this hyperplane to perform classification on the given data. SVM marks training data as points in space, each belonging to one of many classes, in such a way that the categories are divided clearly by a gap that is as wide as possible.[5] It can perform linear as well as non-linear classification. An example of SVM is given in figure 3. In the figure, blue circles represent one class and the red squares represent the other. Both these classes are separated by a hyperplane. The hyperplane is chosen in such a way that it lies equidistant to the closest point of each class (darkened) while the margin between the two closest points of the classes is maximum.

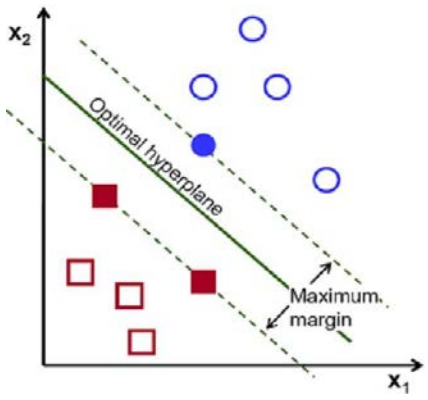


Figure 3: Example of SVM

4.2 Naïve Bayes

Naïve Bayes is a classification technique making use of Bayes' Theorem. In layman terms, it assumes that features of a class don't affect any other feature of the same class. Naïve Bayes is pretty easy to understand and implement. It is particularly useful for large datasets. It is known to perform far better than even the most highly sophisticated classification methods.[6] Bayes theorem is mathematically represented by the equation given below:

$$P(Z|Y) = \frac{P(Y|Z)P(Z)}{P(Y)}$$

The terms used in the formula above are defined below:

Z and Y: Two events

P(Z): Probability of event Z happening

P(Y): Probability of event Y happening (P(Y)≠0)

P(Y|Z): Probability of event Y happening given that event Z happens

P(Z|Y): Probability of event Z happening given that event Y happens

4.3 SVM vs Naïve Bayes

Three datasets of sizes 1,600, 10,000 and 25,000 reviews were used to check the accuracy of the classifiers. After classifying the reviews with both the classifiers, it was found that SVM had a slightly higher accuracy than Naïve Bayes on all the three datasets. The results are presented in a tabular form below:

Size of dataset	SVM	Naïve Bayes
1,600 reviews	78.25%	77.75%
10,000 reviews	86.64%	83.07%
25,000 reviews	87.388%	83.431%

Table 3: Experimental Result

The above data is represented in graphical form below:

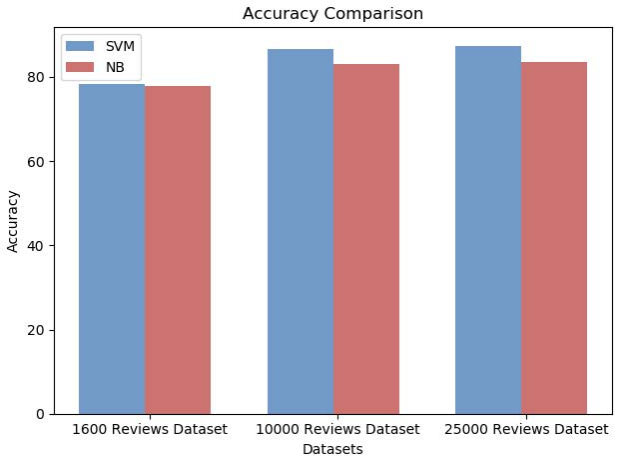


Figure 4: SVM vs Naïve Bayes Accuracy Comparison

From the above figure, we can see that SVM classifies inputs better than Naïve Bayes. The results become more apparent as the training dataset size increases.

5. CONCLUSIONS

The comparison of Support Vector Machine (SVM) and Naïve Bayes methods for binary text classification for sentiment analysis is presented in this paper.

The findings indicate that the SVM classification method for sentiment analysis provides a slightly higher accuracy (min: 78.25%, max: 87.388%) than the Naïve Bayes (min: 77.75%, max: 83.431 %) method.

On further research, it is found that increasing the size of the training dataset from 1,600 to 10,000 and then to 25,000 reviews increases the accuracy of the Naïve Bayes classifier by 5.32% (from 77.75% to 83.07%) when the size of the dataset is increased from 1,600 to 10,000 reviews and by 0.361% (from 83.07% to 83.431%) when the dataset size is increased from 10,000 to 25,000 reviews. The accuracy of the SVM classifier also increases with an increase in dataset size, by 8.39% (from 78.25% to 86.64%) in the first case and by 0.748% (from 86.64% to 87.388%) in the second case.

These results suggest that the accuracy of the classifiers may increase further if a larger training dataset is used.

REFERENCES

[1] Afroze Ibrahim Baqapuri. Twitter Sentiment Analysis (TSA), 2015

[2] Nicholas Diakopoulos, Samuel Brody. Coooooooooooooolllllllllllll!!!!!!!!!!!!!! Using Word Lengthening to Detect Sentiment in Microblogs (CUWLDSM), 2011

[3] Jason Brownlee. Supervised and Unsupervised Machine Learning Algorithms (SUMLA), 2016

[4] Janyce Wiebe, Paul Hoffmann, Theresa Wilson. Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis (RCPPLSA), 2005

[5] Geeks for Geeks. Classifying data using Support Vector Machines (CSVM)

[6] Geeks for Geeks. Naive Bayes Classifiers (NBC)

[7] T Choudhury, V Kumar, D Nigam - An Innovative Smart Soft Computing Methodology towards Disease (Cancer, Heart Disease, Arthritis) Detection in an Earlier Stage and in a Smarter WayInternational Journal of Computer Science and Mobile Computing, 2014

[8] Tomar, Ravi, Prateek, M., & Sastry, G. H. (2016). Vehicular Adhoc Network (VANET) -An Introduction. International Journal of Control Theory and Applications, 9(18), 8883–8888.

[9] K Chandel, V Kunwar, S Sabitha, T Choudhury, S Mukherjee,A comparative study on thyroid disease detection using K-nearest neighbor and Naive Bayes classification techniques ,CSI transactions on ICT 4 (2-4), 313-319

Proceedings of the SMART–2018, IEEE Conference ID: 44078
2018 International Conference on System Modeling & Advancement in Research Trends, 23rd–24th November, 2018
College of Computing Sciences & Information Technology, Teerthanker Mahaveer University, Moradabad, India

Techniques to Elimenate Human Bias in Machine Learning

Eishvak Sengupta[1], Dhruv Garg[2], Tanupriya Choudhury[3] and Archit Aggarwal[4]
[1,2,4]*Amity University, Uttar Pradesh, Noida, India*
[3]*University of Petroleum & Energy Studies, Dehradun, Uttarakhand, India*
E-mail: [1]eishvaksengupta@gmail.com, [2]dhruv98garg@gmail.com,
[3]tanupriya1986@gmail.com, [4]archit.aggarwal1508@gmail.com

Abstract—In an era where human lives have certain dependence on artificial intelligence and machine learning, it is essential for them to make unbiased and accurate predictions. This paper addresses the issue of the inclusion of a human bias in a machine learning algorithm and how it goes to produce skewed results. It goes through the prominent types of human biases and real life incidents where the inclusion of a human bias has had a negative impact. This paper provides a comprehensive review of the methods that can be incorporated to eliminate a human bias focusing on the use of machine ethics making mention of community groups working towards the same.

Keywords: Artificial Intelligence, Machine Learning, Human Bias, Skewed results, Bias dataset

I. Introduction

Artificial Intelligence and machine learning have entered almost all spheres of our lives. From traditional robotics to cybersecurity to something like wildlife preservation, artificial intelligence and machine learning is everywhere. This has led to a point of having a certain degree of dependence of human livelihood on the solutions predicted by these. A machine learning algorithm produces skewed results when either the algorithm is biased or the dataset is predisposed. A data set may be defined as a gathering of related, discrete things of comparable information that was collected via various mediums.

A data set is basically orchestrated into some sort of data structure. In a database, for instance, a data set is a collection of similar data like a classroom in a school. The database itself can be considered a data set, as can groups of information inside it identified with a specific kind of data, for example, deals information for a specific corporate office. The accuracy and precision of the prediction and solution of the model is entirely dependant on the training dataset. A vast dataset would include a greater number of entries and examples exposing the algorithm to greater possibilities and making it more accurate. But the vastness of the dataset does not only refer to a greater number of entries but also a well-distributed one to eliminate any possibility of a bias. The randomness of the entries of the dataset is of great importance so as to avoid skewed predictions. It is essential for both machines and humans to avoid bias in order to prevent any form of discrimination. This paper discusses the certain types of human biases arising due to bias datasets and how the can be eliminated.

II. Machine Learning

Machine learning is a method of teaching computers and programs to make predictions based on some data. It is a branch of artificial intelligence that enables the program to improve automatically. It is a set of algorithms that gives software and applications the ability to end up more precise in foreseeing results without being expressly programmed. Machine Learning in the most elementary sense is the practice of using algorithms to analyze information, get the required data, and afterward make predictions about the real world. The primary aim of machine learning is to create algorithms that receive input data in the form of training sets and use methods of statistical analysis of the training dataset to predict an output while updating the predicted output if any new data is made available in the training dataset. The learning process begins with insights of information, for example, models, coordinate understanding, or guidance, with the end goal to search for patterns in information and settle on better choices later on dependent on the precedents that are given. The basic goal is to let the machines learn automatically without any form of human intervention or assistance and adjust actions accordingly. The more training data is fed into the computer, the better is the performance of its algorithm. Machine learning is used to design algorithms based on the data trends and historical relationships between data. These algorithms allow engineers and data scientists to generate reliable and valid results and decisions enabling some hidden patterns through historical trends in the data. They work on the principle of analysing previous results and making accurate decisions and predictions. Machine learning is most suitable for cases where the theoretical knowledge is incomplete but a sufficient amount of observations and results are present [1].

III. Inclusion of Human Bias

Machine learning can be approached in two ways i.e supervised and unsupervised methods.

In the supervised approach, a database containing a large variety of entries in an attempt to include all possible cases are stored. A machine learning algorithm goes through the entire set of entries and draws up a standard. A supervised learning algorithm takes a known arrangement of information and known response to the information and prepares a model to create sensible expectations for the response to new data that arrives in the form of test cases.

The aim of ML algorithms is to find a predictive model that best generalizes to a particular type of data [2]. Algorithms of Machine learning require a vast number of training datasets that help the algorithm to learn about the system's responses and behaviour so as to predict solutions when new problems are presented to it. The algorithm studies a pattern in the data provided in order to reach a generalized solution for similar types of data.

The dataset for a particular algorithm is prepared by the method of Data Pre-processing. It is the process of converting data with elements of noise from a particular database and formatting it to give it shape such as assigning proper columns for the nameless or missing feature names. The datasets arriving sometimes have entries that are invalid, not present or otherwise in a format that is difficult to be processed by the algorithm. In the case of invalid data the algorithm is not able to function to its full potential and results in data of lower accuracy with misleading consequences in some cases. Good data preparation produces spotless and well-gathered data that leads to accurate outcomes. It also includes converting the data into information which is relatively easier to understand. The datasets provided to the algorithms for learning are prepared by humans (data analysts). A bias is any disproportionate inclination towards or against any idea, individual or belief. As the machines learn from these human-provided examples rather than explicit rules, programming and constraints, the bias of the human, in most cases unintentionally, is bound to creep into the dataset making the results skewed toward a certain aspect of the data. Biases find their way into the systems mostly through data and fewer times through an algorithm. Biases which are developed on data related to human entities often has a tendency to resemble human-like biases towards race, sex, religion, and many other common forms of discrimination [3]. As a result, when the algorithm begins to function along with the human bias, it gets converted into a large-scale bias. The decisions made by machines have a significant effect on people's lives. These algorithms have repeatedly shown their shortcomings due to the inclusion of a human bias. These are known to have direct and indirect consequences in the livelihoods of people. Generally speaking, an algorithm cannot be biased as such because ultimately it is the output of the statistical analysis of data provided. Hence the onus of accurate results lies on the data. Robotized innovations are produced by people, so that our human biases aren't able to enter the software and systems to make things simpler. But due to their failure to reason for the human bias, the results may be skewed and holds the potential to harm livelihoods of minority groups of the respective field for which the systems are built.

IV. Types of Human Bias

Artificial intelligence is subject to cognitive bias, just like the human brain. Human cognitive biases are processes that disrupt decisively and reasoning ability, ending up in errors. Human bias instances include stereotyping, the bandwagon effect, affirmation predisposition, priming, selective perception, the speculator's false notion, and the observational selection bias. The total number of cognitive biases is constantly evolving, due to the ongoing identification of new biases. A human bias is a type of cognitive bias in which one variant is given significant preference over the other. In machine learning, a human bias gets incorporated not because of a defect in the algorithm but because of the presence of a bias in the training dataset. Human bias in machine learning are of different kinds. Around 180 human biases have been identified. Some of the most prominent ones are:

A. Interaction Bias

When a machine is fed a dataset containing entries of one particular type, an interaction bias is introduced which prevent the algorithm from recognizing any other types of entries. In an algorithm made to identify phones, if a great majority of the entries are large display touch screen phones the algorithm fails to recognize button type phones whose entries in the dataset are comparatively very less. The algorithm makes a generalised prediction assuming all phones have a large display and don't have buttons.Here the feature of a large screen and no buttons is dominating over button type phones only because Hence the algorithm displays a bias against button type phones.

B. Latent Bias

A latent bias is experienced when multiple examples in the training set have a stand out common characteristic, the ones without that characteristic are failed to be recognized by the algorithm. A research conducted at the MIT Media Labs on the accuracy of a facial detection software demonstrated a clear example of a latent bias. 1270 faces were included into the dataset, using the faces of politicians, which included a strong percentage of women holding positions in public offices. Facial recognition software developed by IBM, Microsoft and other similar Chinese companies were being tested for their accuracy. The software had shortcomings in identifying faces dependent on a person's skin colour. The algorithm was biased towards white people and failed to identify dark-skinned faces. The algorithm was further biased towards women [4].

C. Selection Bias

Selection bias is introduced to an algorithm by the selection of data for analysis in such a way that proper randomization is not achieved. When an algorithm is made

for the purpose of identification of faces, the dataset cannot possibly include all types of facial structures and shapes. The entries in the dataset will be highly dependant on ethnical and geographical constraints. Even if images are incorporated from the internet, it is seemingly impossible to include every face type reducing the randomness of the dataset thereby introducing a selection bias.

V. Incidents of Negative Impact of Human Bias in Machine Learning Applications

A. Tay-the Chatbot

In 2016, Microsoft released its Twitter chatbot named Tay. Tay was an early attempt by Microsoft of incorporating Artificial Intelligence and Machine Learning in its products full-fledged. The bot was made such that it would blend in with millennials and talk like them. But in less than 24 hours, Tay began making racist Nazi comments like "Hitler was right". The Machine learning algorithms of Tay were made such that their training datasets were made up of the activity of other accounts interacting with it i.e it used to learn by reading and understanding other tweets and comments [5]. I was expected that Tay to get better at responding and answering to tweets as more people engaged with her. Tay was an ideal A.I. chat bot and was supposed to exhibit one of the most important features of true A.I.—the ability to get smarter, more effective, and more helpful over time. Its algorithm had the ability to analyse data more quickly, and in a more accurate manner. However, it inherited human biases and prejudices. Because of a bias in the accounts that Tay interacted with, the results produced by it were skewed toward a particular community with whom the algorithm learnt mostly from. The chatbot was removed within a day owing to an extremely biased learning dataset. This was an example of how a human bias was getting converted into a large scale bias creating negative impact.

B. Speech Recognition Software

Other places where human bias can be encountered is speech recognition software. Natural language processing (NLP) and deep learning neural networks are the drivers of a speech recognition software. The software disintegrates the speech into smaller fragments that can be interpreted into a digital format, and can analyse the fragments. After trying to determine the verbal contents of the user's speech based on code and other criteria, the system dictates the results into text. Speech recognition completely works on computational linguistics.

The speech recognition system (SRS) is basically a pattern recognition system, including feature extraction, pattern matching, the reference model library [6]. But a constraint in the model library, owing to a less expansive dataset, the algorithm is being introduced to produces biased results. In order to train SRS, designers utilize extensive datasets, which may be recorded individually, or given by other semantic specialists. Also, in some cases, these datasets do exclude assorted speakers. An American voice recognition software fails to understand the Scottish accent. The challenge for AI is in programming a changing vocabulary related semantic into a binary numerical system. The challenge for Machine learning is in programming a changing vocabular or ethnic or religious compositions into a binary numerical system. The rate of error detection in local dialect and genders is differential, and that making variations in pitch would not be sufficient to make the recognition system less responsive for that speaker. While the last needs extra information to shape a strong speculation, the size of the effect for the former is deeply disturbing. From a linguistics point of view, no vernacular is characteristically pretty much understandable.

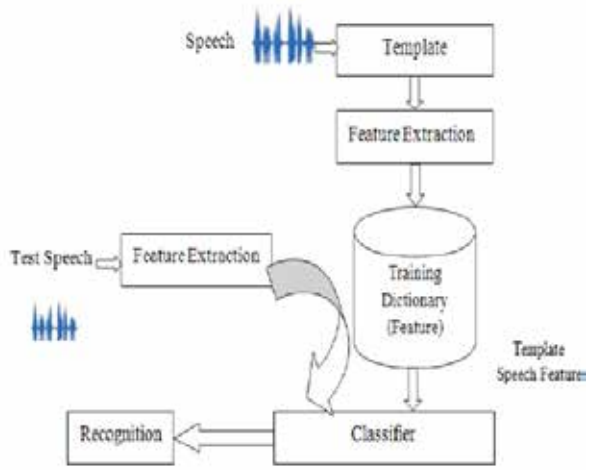


Fig. 1: Source: Https://www.researchgate.net/Figure/Working-of-Speech-Recognition-Process_fig1_281684338

Human intervention is extremely necessary to adjudicate the bias in the programmer, the context and the language itself. Rachael Tatman, linguist researcher, determined Google's speech recognition software has encountered bias based on gender. The main discovery was made on the auto caption system of YouTube where the voices of male figures were predicted better than female voices by a staggering amount. She said the outcomes were "profoundly exasperating." Tatman said she hand-checked more than 1,500 words from annotations across 50 different videos and discovered a glaring bias. The rate of error detection in local dialect and genders is differential, and that making variations in pitch would not be sufficient to make the recognition system less responsive for that speaker. While the last needs extra information to shape a strong speculation, the size of the effect for the former is deeply disturbing. From a linguistics point of view, no vernacular is characteristically pretty much understandable [7]. Gender bias is not just in the algorithms. It lies within the outcomes–predictions and recommend. AI assistants powered mainly by female voices and personas on all kinds of devices (examples include Alexa and Siri), are mostly seen as helpful or passive supporters of a user's lifestyle

which are considered to be inherent features of women. Whereas, male equivalent assistants in the likes of IBM's Watson or Salesforce's Einstein are perceived as complex problem-solvers tackling global issues. One method to re-advocate this perception is to turn such assistants genderless.

Mozilla's voice assistant common voice uses a slightly different approach—it uses an open source dataset. Mozilla put out a call to gather different accents of voices and has been able to gather thousands of such datasets in a wide range of languages, English being the primary. Anyone in the world poses the ability to record the pre-determined sentence. These voices are then used to train Mozilla's own algorithm known as Deep Speech.

C. Law Enforcement using Machine Learning

In the recent past, law enforcement has been one of the major application of artificial intelligence. One such application is predictive policing, which utilises certain algorithms in AI to determine the stats like the location and nature of the crimes which may happen in the future. The system determines patterns and combinations based on past conviction records and then feeds these patterns to a predictive model, where they are combined in order to calculate the probability for an individual whose future is as-yet unknown.

Machine learning algorithms consequently distinguish beforehand obscure pictures and video cuts identified with key classes, for example, child abuse, weapons, cash, drugs, nakedness and more to instantly pinpoint comparable things, for example, faces, articles, images or topics sparing valuable time.

a machine learning algorithm, built on a so-called "partially generative neural network." The algorithm takes input data such as the racial, sexual and religious composition of a localised area and predicts the occurrence of the criminal activity [8]. PredPol, a US based predictive policing software, uses a machine-learning algorithm to calculate its predictions. The algorithm learns from past data of about 2 to 5 years about each new city it is being introduced to. It makes use of three data points–crime type, location, and date/time–to create its predictions. It Predicts where and when specific crimes are most likely to happen. Initially, the model generated race and religion biased predictions. This was a direct outcome of a skewed dataset against a particular race and religion. Kristian Lum, an American researcher, stated that the model was reinforcing the general biased mindset of the police department and was hence producing biased results [9].

D. Advertisements and Autofill

A similar problem is noted to biases in algorithms that determine content on the web for instance, criminal records turn up higher in searches for names commonly associated with someone of African-American descent , women get served ads for lower-paying jobs, and the software some law enforcement agencies use is biased against people of colour. The number of adverts for arrest records were altogether more inclined to appear on scans for unmistakably black names or possessing any sort of link to black ethnicities [10]. The algorithm in all probability did this for everyone, but over the course of time the biases of the people who did the search got factored in. algorithms and online results simply reflect people's attitudes and behavior. Machine learning algorithms learn and evolve studying a user's activity online. The autocomplete feature used in search engines is an example. A recent Google search for "Are transgender," for instance, suggested, "Are transgenders going to hell."

VI. Eliminating Human Biases

Identifying and recognising biases is the most challenging problem, and research has been done in order to prepare methods to alleviate these biases. The prepared models are trained on data sets that may not adequately speak to an objective populace.

The use of machine ethics should be strongly emphasised to intercept and amend for any biases in machine learning algorithms. The moral aspect machine learning is overviewed by machine ethics. Machine ethics are not the same as robo-ethics, which deals with the morality of engineers working towards creating and the working of robots. Machine ethics also differ from computer ethics, which focuses on data security. The three essential methods of incorporating ethics so as to utilize them to alleviate negative injustice in algorithmic programming are technical, political, and social. Finding comprehensive data, experimenting with different datasets and metrics, increased representation in the technical workforce, external validity testing, and auditing are some adoptable methods to remove human bias factor. When preserved characteristics such as age, gender, and race are parameters of an algorithm, it is essential to join them while additionally tending to the social bias that accompanies from a particular attribute within the code [11]. Community groups such as the Algorithmic Justice league founded by MIT Media Labs post the failure of the facial recognition algorithm, help to advance publicly supported announcing and the study of bias in machine learning and artificial intelligence. Inclusion from various populaces in the moral creation and consumption of machine learning predictions will lead to further progress in ethics that include all users. The problem of bias can also be resolved in intelligent systems using techniques of emotion recognition. Most emotion recognition techniques make use of images recognition algorithms. Using a cloud-based emotion recognition algorithm applied to images associated with a minority class can help remove and skewness towards a minority [12].

Using deep learning algorithms will significantly help to reduce the inclusion of any sort of bias in the algorithm.

Deep learning is a sort of a super set of machine learning methods based on learning data representations, whereas there are more issue specific algorithms in machine learning. Deep-learning algorithms are seeing rapid growth in terms of usage especially in corporate sectors for screening and evaluation of employees and in law departments for various purposes

Incorporating methods like bias testing will help represent the vulnerable sides made due to the lack of a wide variety of people who build these systems. It would help discuss issues which might not be obvious to the data analyst and, most importantly, to the end user. For instance, how a virtual assistant should respond when repeatedly questioned. A certain level of bias testing will have to continue after a prediction is made, since the algorithms continuously evolve

The IBM- MIT Watson lab making optimal use of recent developments in the field of artificial intelligence and computational cognitive modelling in the form of contractual approaches to ethics with the aim to shape technologies that apply certain human values and principles in decision-making [13].

As AI based technologies continue to integrate into the daily schedules of people around the world for constructive purposes not only in the form of voice assistants like Amazon's Alexa or Google Home, AI-driven industrial technologies will help to enhance overall productivity, close workforce aptitude holes and support client encounter crosswise over enterprises. This is the most opportune time to adopt methodologies to eliminate human biases and remove any parameter of gender, include more engineers and analysts working on these technologies, and address trust issues with AI.

VII. Conclusion and Further Scope

Although the presented approach demonstrates legitimacy in tending to the issue of bias, there are still a number of exception cases of biases that have scope of research. Future work in this domain includes determining an approach with a focus on a different minorities and validating them with different generalized machine learning algorithms. This paper was written with the purpose to add to raising the affectability to the potential difficulties in the execution and bias of classifiers when making inferences about people belonging to different groups and skin colours.

References

[1] David J.LaryaAmir H.AlavibAmir H.GandomicAnnette L.Walkerd "Machine learning in geosciences and remote sensing" Geoscience Frontiers Volume 7, Issue 1, January 2016, Pages 3-10

[2] Tanushri Chakravorty "How Machine Learning Works: An Overview" 2016 https://thenewstack.io/how-machine-learning-works-an-overview/

[3] Daniel James Fuchs Missouri University of Science and Technology May 2018"The Dangers of Human-Like Bias in Machine Learning Algorithms"

[4] Joy Buolamwini, MIT Media Labs, "Facial recognition software is biased towards white men, researcher finds" Feb. 11, 2018, retrieved 30th September 2018

[5] Albayrak, N., Ozdemir, A., & Zeydan, E. (2018). An overview of artificial intelligence based chatbots and an example chatbot application. 2018 26th Signal Processing and Communications Applications Conference (SIU).

[6] Meng, J., Zhang, J., & Zhao, H. (2012). Overview of the Speech Recognition Technology. 2012 Fourth International Conference on Computational and Information Sciences.

[7] Rachael Tatman Department of Lingusitics University of Washington "Gender and Dialect Bias in YouTube's Automatic Captions"

[8] Buscema, Paolo Massimo, Tastle, William J. (Eds.) 2013 "Intelligent Data Mining in Law Enforcement Analytics New Neural Networks Applied to Real Problems"(Springer)

[9] Kristian Lum, "Predictive Policing Reinforces Police Bias" OCTOBER 10, 2016

[10] Latanya Sweeney, Harvard University "Discrimination in Online Ad Delivery" January 28, 2013

[11] Ellis, Geoffrey (Ed.)2018 "Cognitive Biases in Visualizations"(Springer)

[12] Howard, A., Zhang, C., & Horvitz, E. (2017). Addressing bias in machine learning algorithms: A pilot study on emotion recognition for intelligent systems. 2017 IEEE Workshop on Advanced Robotics and Its Social Impacts (ARSO).

[13] IBM research IBM-MIT Watson labs blog.