Criterion 3: Research, Innovations and Extension

3.2.1 Number of books and chapters in edited volumes/books published and papers published in national/international conference proceedings per teacher

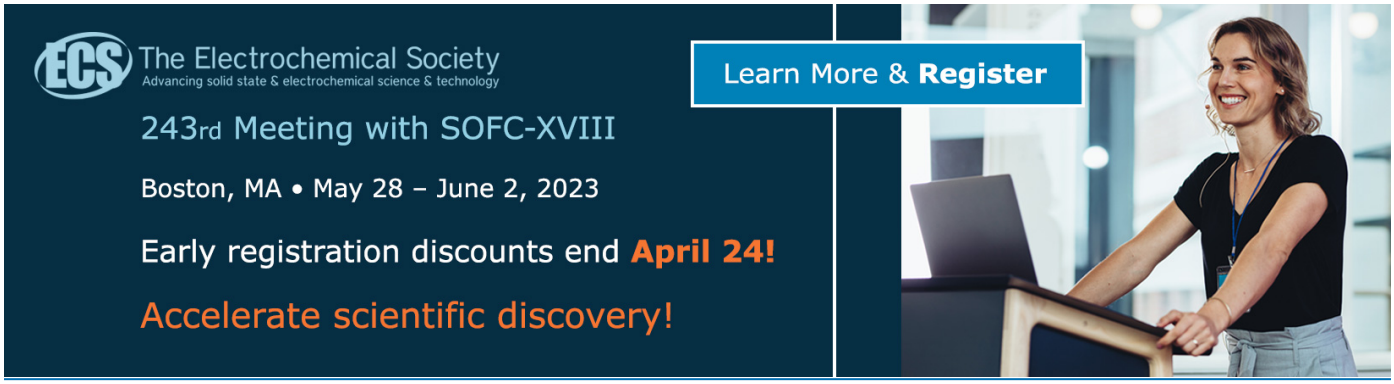
AMBILY MERLIN KURUVILLA

Certificate of Presentation

Proceedings

Heart disease prediction system using Correlation Based Feature Selection with Multilayer Perceptron approach.

To cite this article: Ambily Merlin Kuruvilla and N.V Balaji 2021 *IOP Conf. Ser.: Mater. Sci. Eng.* **1085** 012028



Heart disease prediction system using Correlation Based Feature Selection with Multilayer Perceptron approach.

Ambily Merlin Kuruvilla[1] **and Dr.N.VBalaji**[2]

[1] Research Scholar, Department of CS,CA and IT, Karpagam Academy of Higher Education
[2]Dean, Faculty of Arts, Science and Humanitics, Karpagam Academy of Higher Education

E-mail: E-mail:ambilykuruvilla@gmail.com

Abstract. Cardiac disease prediction helps physicians to make accurate recommendations on the treatment of the patients. The use of machine learning (ML) is one of the solution for recognising heart disease-related symptoms. The goal of this study is to suggest a methodology for identifying the most relevant features of cardiac disease characteristics by applying a feature selection technique. The data set used in this study was Framingham heart disease dataset (FHS). It was collected from KAGGLE Machine Learning repository. There are 16 attributes and a mark in the dataset that has been validated by four ML classifiers. There are two feature selection methods, Correlation Based Feature selection (CBFS) and Principle Component Analysis (PCA) was used for the comparison in the study. By using CBFS Method five highly correlated features are selected for the study, and by using PCA thirteen features are selected. The experimental result shows that Correlation Based Feature Selection with Multilayer perceptron (CBFS with MLP) obtained the highest accuracy for this dataset.

1. Introduction

The research concentrates on the two feature selection methods for data reduction before building the predictive models by classification algorithms. These reduced features are then passed into the classification algorithms to design the models for the heart disease prediction. These models are used for the comparison of accuracy of the classifier. Principle Component Analysis and Correlation Based feature selection methods are used for finding out the reduced features. The selected features are inputted to four different classifiers such as Navie Bayes, ADABOOST, MLP and SMO. The accuracy of each model is compared with the other.

2. Background Study

Devansh Shah studied various attributes related to heart disease[1]. The study was conducted with Naïve Bayes, decision tree, K-nearest neighbor, and random forest algorithms[1]. The experimental result proves that K-nearest neighbor algorithm exhibits the highest accuracy.

Hamidreza Ashrafi Esfahani[2] formulated a model to predict cardiovascular disease. The model includes decision trees, Neural Networks, Rough set, Naïve Bayes and SVM for implementation. On comparing the results achieved, it was revealed that the hybrid model of Rough Set, Naïve Bayes and Neural Network obtained the highest accuracy. An ensemble strategy was implemented that allowed for the output to be combined that would result in

better accuracy. The performance of the classifiers was compared with the parameters such as its precision, sensitivity, accuracy and F-Measure.

3. Proposed Methodology

Framingham Heart Study (FHS) from Kaggle Machine Learning repository is used for the study. Two feature selection methods along with four classification algorithms are used for the study. CBFS and PCA are the methods used for the dimensionality reduction. MLP, Navie Bayes, Sequential Minimum Optimiser (SMO) and ADABOOST algorithms are used for classification. The reduced feature set from both the feature selection methods are inputted to different classifiers. Eight different Machine Learning Models were created for Heart Disease Prediction. Accuracy of these Models are compared with each other.

4. Description of the Data Set

The Framingham Heart Study (FHS) dataset was collected from Kaggle. The dataset consists of 4241 records. It contain sixteen features including AGE, PREVALENT HYP, SYSBP, DIABP, GLUCOSE, SEX, EDUCATION, CURRENT SMOKER, CIGSPERDAY, BPMEDS, PREVALENT STROKE,BMI, HEART RATE, DIABETES, TOTCHOL and PREDICTOR VARIABLE.

5. Classification Algorithms

In Machine Learning various forms of classification techniques are available. Classification techniques used for this study was described below.

6. Multilayer Perceptron (MLP)

MLP is a subset of Artificial Neural Network. MLP comprises one or more than one hidden layers aside from one input and one output plate. The Perceptron is made of an input layer and a totally linked output layer. MLPs have the same levels of input and output, but could have several levels concealed within them.

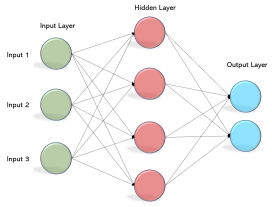


Figure 1. Different layers in MLP.

7. Adaboost

In machine learning, AdaBoost(Adaptive Boosting) is a supervised learning algorithm. It is used for combining several weak classifiers together to generate a strong classifier.

8. Naive Bayes

Naive Bayes is a Machine Learning algorithm based on Probability theory in statistics. The term naive suggests that the elements that go through the software are autonomous of each other,That is, the value of one characteristic, does not explicitly influence or alter the value of any of the other characteristics used in the algorithm. The Bayes theorem tells us how we can compute the conditional probability. The equation for conditional probability is,

P(Ai/Bi) =(P(Bi/Ai) * P(Ai)) /P(Bi)

P(Ai/Bi) defines the probability of an event Ai occurs corresponding to the event Bi has occurred.

P (Bi / Ai) is the conditional probability and it defines the probability of occurrence of event Bi corresponding to the occurrence of the event Ai.

P(Ai) and P(Bi) defines the probability of the events Ai and Bi occurs.

9. Sequential Minimal Optimization (SMO)

The sequential minimal optimization is more effective to solve the SVM problem compared to traditional Quadratic Programming algorithms such as the interior-point method. The SMO algorithm can be viewed as a method of decomposition by which a problem of optimization of multiple variables is decomposed into a set of sub problems, each optimizing an objective feature of a limited number of variables, usually only one, whereas all other variables are treated as constants which remain unchanged in the sub problem.

10. Feature Selection

During feature selection the most relevant features are extracted from the data set. Redundancy can be avoided using this method. Since irrelevant features are excluded from the input data, feature selection can increase the accuracy of prediction. In this study Correlation Based Feature Selection (CBFS) and Principle Component Analysis(PCA) is used for feature selection. After feature selection the reduced data set is applied to four different classification Algorithm.

11. Correlation Based Feature Selection(CBFS)

Correlation values are calculated by CBFS. The five highly correlated features are selected for the analysis. These features are given as the inputs for the classifiers.

Table 1. Features selected for the analysis by CBFS along with the correlation values.

S/N	Selected Features	Correlated Values
1	AGE	0.2254
2	PREVALENTHYP	0.2164
3	AGE	0.2254
4	PREVALENTHYP	0.2164
5	PREVALENTHYP	0.2164

12. Principal Component Analysis (PCA)

Thirteen features were selected by PCA during feature selection. The features selected by the PCA algorithm are AGE, PREVALENTHYP, SYSBP, DIABP, DIABETES, SEX, BPMEDS, TOTCHOL, PREVALENTSTROKE, CIGSPERDAY, EDUCATION, BMI, CURRENT SMOKER.

13. Result and Discussion

In the study two feature selection methods are used for comparison - Principle Component Analysis (PCA) and Correlation Based Feature Selection (CBFS). After dimensionality reduction the reduced dataset is applied to four different classification Algorithm such as

Multilayer Perceptron (MLP), AdaBoost, Navie Bayes and SMO. Five most correlated features were selected and applied to Classifiers in CBFS Method. Thirteen features were selected by Principle Component Analysis (PCA). The result is shown in Table4. From the results it is found out that Correlation Based Feature Selection (CBFS) along with MLP algorithm shows maximum accuracy.

Table 2. Comparison of Predictive accuracy of Models.

S/N	Alogrithm	CBFS	PCA	Before FS
1	MLP	84.9057	83.9151	84.1509
2	ADABOOST	84.8113	84.8821	84.8821
3	Navie Bayes	81.1792	80.0472	80.0472
4	SMO	84.8113	84.8113	84.8113

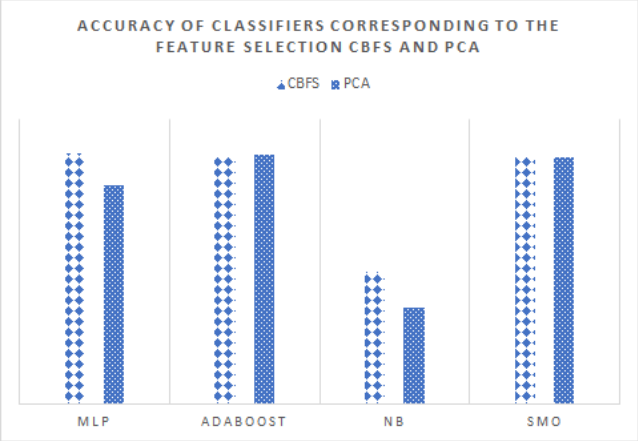


Figure 2. Comparison of Accuracy of classifiers corresponding to CBFS and PCA Feature Selection Methods.

Eight Models are generated by combining two feature selection algorithms and four classifiers. They are CBFS-MLP, CBFS-ADABOOST, CBFS-NB, CBFS-SMO,PCA-MLP, PCA-ADABOOST, PCA-NB and PCA-SMO.The accuracy of various Models are shown in Table.1 The results from table proves that Correlation Based Feature Selection along with Multilayer Perceptron (CBFS-MLP) Model perform better than the other Models with the accuracy of 84.9057 Percentage.

The result from Table 2 proves that CBFS-MLP combination shows better performance than MLP classifier. Also when we are comparing CBFS-MLP and PCA-MLP models accuracy measures proves that CBFS-MLP combination shows more performance.

Table 3. Comparison of Predictive Accuracy of MLP.

MLP	Precision	Recall	F Measure	ROC
Before FS	0.790	0.842	0.798	0.671
After CBFS	0.805	0.849	0.791	0.649
After PCA	0.784	0.795	0.668	3.62

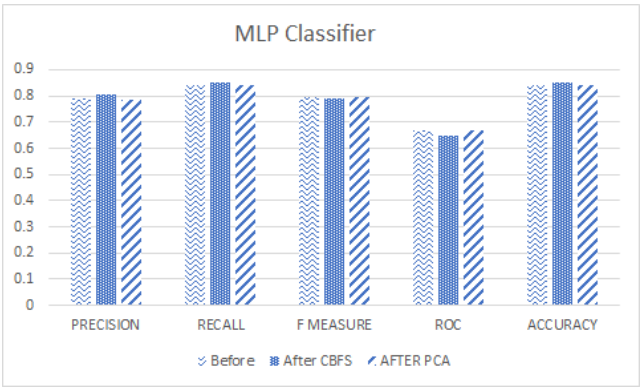


Figure 3. Comparison of Accuracy of MLP Classifier.

14. Conclusion and Findings

During the study the performance of two different feature selection methods CBFS and PCA are evaluated. Eight different classifier models are developed by combining the feature selection and classification algorithms.The performance of each model was evaluated.Performance measures such as Accuracy, Precision, Recall, F Measure and ROC are evaluated for finding out the best classifier. From the result it is proven that the model CBFS with MLP Classifier shows the maximum performance for FHS dataset.

References

[1] Devansh Shah, Samir Patel and Santosh Kumar Bharti 2020 Heart Disease Prediction using Machine Learning Techniques Springer.
[2] Hamidreza and Ashrafi Esfahani 2017 Cardiovascular disease detection using a new ensemble classifier IEEE International Conference on Knowledge-Based Engineering and Innovation (KBEI).
[3] V.V.Ramalingam 2018 Prediction of Heart Diseases Using Machine Learning International Journal of Engineering and Technology.
[4] Amin Ul Haq 2019 A Hybrid Intelligence System Frame Work for Prediction of Heart Disease Using ML, Mobile Information Systems.
[5] Shadman Nashif Heart Disease Detection by Using Machine Learning Algorithm and a Real time Cardiovascular Health Monitoring System World journal of Engineering and Technology,vol.06 no.04,2018,article id 88650.
[6] Poornima Singh 2018 Effective heart disease prediction system using data mining techniques International journal of Nano medicine.
[7] M. Gunay and T. Ensarı 2019 Predictive churn analysis with machine learning methods IEEE 26th Signal Processing and Communications Applications Conference (SIU), Izmir, 2018, pp. 1- 4.
[8] R. Suguna, M. Shyamala Devi and Rincy Merlin Mathew 2019 Customer Churn Predictive Analysis by Component Minimization using Machine Learning International Journal of Innovative Technology and Exploring Engineering.
[9] Shyamala Devi, Munisamy, Suguna Ramadass and Aparna Joshi 2020 Cultivar Prediction of Target Consumer

Class using Feature Selection with Machine Learning Classification Springer's book series "Learning and Analytics in Intelligent Systems Springer, LAIS vol. 3, pp. 604-612, 2019.
[10] Suguna Ramadass and Shyamala Devi 2019 Prediction of Customer Attrition using Feature Extraction Techniques and its Performance Assessment through dissimilar Classifiers Springer's book series Learning and Analytics in Intelligent Systems, Springer.
[11] R.Suguna, M. Shyamala Devi, Rupali Amit Bagate and Aparna Shashikant Joshi 2019 Assessment of Feature Selection for Student Academic Performance through Machine Learning Classification Journal of Statistics and Management Systems, Taylor Francis.
[12] M. Shyamala Devi, Rincy Merlin Mathew and R. Suguna 2019 Feature Snatching and Performance Analysis for Connoting the Admittance Likelihood of student using Principal Component Analysis International Journal of Recent Technology and Engineering.
[13] Debjani Panda, Ratula Ray, Azian Azamimi Abdullah and Satya Ranjan Dash 2019 Role of Feature Selection in Prediction of Heart Disease International Conference on Biomedical Engineering (ICoBE) IOP Publishing
[14] Bandari Sai Santosh, Dharma Sahith Reddy, M Sai Vardhan and Dr. Shaik Subhani 2019 Heart Disease Prediction with PCA and SRP International Journal of Engineering and Advanced Technology .
[15] Mothe Sunil Goud 2019 Heart Disease Prediction and Performance Assessment through Attribute Element Diminution using Machine Learning International Journal of Innovative Technology and Exploring Engineering.

Certificate of Presentation

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 8, Special Issue 1, August 2021

HYPER PARAMETER OPTIMIZATION IN STACKED DEEP NEURAL NETWORK FOR MEDICAL DIAGNOSIS

Ambily Merlin Kuruvilla
Research Scholar
Department of Computer Applications, Karpagam Academy of Higher Education, Coimbatore, India.
ambilykuruvilla@gmail.com

Dr. N V Balaji
Dean, Faculty of Arts, Science and Humanities, Karpagam Academy of Higher Education
Coimbatore, India.
balajinv@karpagam.com

Abstract— Diabetes is a metabolic disease where the blood sugar rate of an individual is consistently above normal. Due to the modern lifestyle and work culture, diabetics is widespread and affects the productivity and quality of life for an individual. A diabetics patient is at a very high risk of various health issues like organ failure and even it can even result in loss of life. An early prediction of this chronic disease can avoid health issues and save many lives. The aim of this article is to develop a better predictive model for diabetics using an automated hyper parameter optimization (HPO) approach in Multilayer Perceptron (MLP). This article provides an efficient way to increase the accuracy of Neural Network to a substantial level through the HPO process using Grid Search Optimization (GSO) through the stacking ensemble model. In order to run the ensemble model at an optimal level and to minimize errors, appropriate hyperparameters must be calculated. Three GSO methods are utilized to tune the hyper parameters. To build the stacking ensemble model, the PIMA data set was used.

Keywords: HPO, MLP, GSO: Hyper Parameter Optimization, Multilayer Perceptron, Grid Search Optimization.

INTRODUCTION

A study carried out by the WHO recently revealed that in 2016, diabetics was one of the leading causes of death worldwide. Diabetes has resulted in 1.6 million fatalities in 2016 and this statistic replaces HIV / AIDS with diabetes as one of the most frequent cause of death [4]. The burden of diabetes disease grew from 108 million in 1980 to 422 million [5] in 2014, and the percentage of diabetic patients amongst adults over 18 years of age rose from 4.7% in 1980 to 8.5% in 2014[5]. 642 million people i.e. (1 in 10 people) are expected to contract diabetes by 2040. 46.5% of people with diabetes have not been diagnosed officially [6]. This makes it necessary to develop techniques and procedures to assist in the early detection of diabetes in order to reduce the number of deaths related to diabetes, as late diagnosis is responsible for a majority of deaths linked to diabetes [7].

There is a need to implement sophisticated information processing to develop cutting-edge strategies for the early detection of diabetes. Data mining tools can also be effectively applied. The ability to remove and uncover previously unseen, secret, yet important patterns from a large database repository is given by data mining [7]. These tools can assist medical evaluation and decision making.

LITERATURE SURVEY

Roshan Birjais [1] conducted a research in many classification algorithms for diabetes prediction and his team found out that the Gradient Boosting algorithm outperform other classifiers and

obtained 86% accuracy. They have analysed various prime factors for the cause of diabetes disease.

Md. Maniruzzaman [2] wrote an article in which he used LR for identifying the prime features for the diagnosis of diabetes disease. Accuracy and Area Under curve are used as the performance measures for the classifiers. It revealed that LR along with the Random forest generates the maximum accuracy.

Atik Mahabub[3] used an ensemble voting classifier for forecasting diabetes disease. The Out of eleven classifiers, the best performing three will be used for the ensemble classifier. Accuracy, Precision, F-Measure and Recall [3] are used as the parameters for evaluating the classifiers. The outcome clearly shows that the ensemble method outperform the base classifiers.

BACKGROUND STUDY

Neural network is used for the prediction. Parameter tuning mechanism in neural network is applied for improving the accuracy.

ARTIFICIAL NEURAL NETWORK

Artificial Neural Network [17] is built by multiple nodes that reproduce the human brain's biochemical neurons. The neurons are linked and are communicating with each other. The nodes are capable of taking input data and executing the operations. The outcome of these operations is transferred to other neurons. The output is referred to as its activation or node value for a node. A technique known as Gradient Descent in the Artificial Neural Network, which takes places in the backpropagation period whereby it intends to regularly resample the gradient of the model parameter in the reverse direction based on the weight 'W', periodically updating till the global minimum of G(W) function is reached.

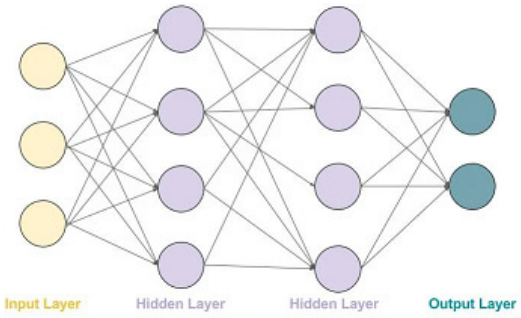


Figure 1: Hidden Layers in Artificial Neural Network

A loss represents the prediction error found in the Artificial Neural Network. In deep learning, this is estimated as a loss function. The Loss Function describes the model's operating efficiency. For the estimation of the loss function, stochastic gradient descent was used. For each iteration, weights are updated, and the model is trying to reach the global minimum point.

LOGISTIC REGRESSION

Logistic Regression is a well-known classification algorithm used to estimate the probability of a target variable. The design of the goal or dependent variable is binary, indicating that only two possible groups are available. Mathematically, f(Xi), a LR model predicts P(Yi=1).

b=wt0+wt1 x1+wt2x2+…+wtnxn.

A sigmoid function can be used to map the predicted values to the probabilities.

S(y)=1/(1+e(-y))

S(y) is the estimated output; y is the value inputted to the function and e is the base of natural log.

NAVIE BAYES

The Navie Bayes algorithm follows the Bayasian principle which belongs to a category of conditional probabilities (CP). The CP is the probability that an event B will happen, provided A has already occurred. Though Bayes Theorem provides a principled way for calculating conditional probability, in practice its computationally expensive and thus using some assumptions, bayes theorem is simplified by making some assumptions and turning it into an effective classification model referred to as Naive Bayes. Conditional probability can yield the probability of an event using prior information.

P(H/E) = (P(E/H) * P(H))/(P(E))

PROPOSED MODEL

Stacking Ensembles

To obtain better performance, ensemble methods allow for combining the results of many methods. More the models, better will be the performance of the ensemble strategies. An ensemble method where the no. of models is stacked in a way that their observations act as input to a new model is called Stacking.

Hyper Parameters in Neural Network

The variables that determines how the network is trained are called the Hyperparameters. Hyper Parameters also determine network structure. These variables are set before the training phase, i.e. before the weight and bias is optimized.

Network Weight Initialization

Based on different activation functions applied to each layer, it is preferred to use separate weight initialization schemes.

Activation function

It is used to apply nonlinearity to frameworks that will permit nonlinear prediction restrictions to be learned by deep learning models. The most popular among the activation functions is the rectifier activation function. While making predictions, for binary, Sigmoid is used while for multi-class predictions, softmax is deployed in the output layer.

Gradient Descent

The learning rate determines how fast its parameters are modified by a network. The learning process is slowed down by a low learning rate, but converges efficiently. The higher learning rate accelerates learning, but does not converge. A decreasing learning rate is usually prioritised.

Momentum

Momentum is used to eliminate oscillations and to know the course of the next step with knowledge of the previous phase. For momentum, usually, a value between 0.5 and 0.9 is used.

Number of epochs

The number of epochs measures the number of times the entire training data is given to the model during the training process. The number of epochs is raised until the accuracy begins to decline, while the accuracy of the training is improved due to overfitting.

Batch size

Refers to the no. of sub samples provided to network before the parameter is updated.

METHODOLOGY

Diabetes prediction using MLP Model is described in the following steps. In this model these steps are used to predict diabetes more accurately.

Data Collection

PIMA Indian Data set [17], from the UCI repository is used for the analysis. The data set consists of features including age, number of pregnancies, Body Mass index, Diabetes Pedigree Function, Glucose Level [17] etc…

HYPER PARAMETER TUNING USING GRID SEARCH.

It is one of the traditional hyper parameter tuning method [13]. Before the learning process starts, the value of the hyper parameter needs to be calculated. Grid Search is also known as a comprehensive search [13], with each mixture of hyper parameters explored by Grid Search. This means that each variation of the hyper parameter values listed would be tried. There can be several parameters for models, and it can be viewed as a search problem to find the right combination values to the parameters. The purpose of algorithm tuning is to find the best values of the parameter corresponding to the particular problem. Grid Search can be expanded to provide the highest results by using automatic approaches for finding optimum values to the parameters.

HYPER PARAMETER TUNING USING GRID SEARCH.

Also known as a comprehensive search, it's a traditional hyper parameter tuning method. The value of the hyper parameter needs to be measured before the start of the learning process. Here, each variation of the hyper parameter values is explored which means that each variation of the hyper parameter values that are listed is tested. On further expanding, grid search can provide the highest results using automatic approaches that would optimum values for the parameters.

Implementation

Input values are converted to vectorised format using Logistic regression. The output of Logistic Regression is saved as a collection in Python and this will be given as the input for the MLP. Grid Search is used for doing the parameter optimization. The model optimizes the following hyper parameters

- Epoch
- Batch Optimization.
- Gradient descent

By using Keras Classifier Grid parameters such as Batch size and Epochs are optimized. In this model stochastic Gradient is introduced for training the Artificial Neural Network. The internal parameters such as epoch and Batch values are optimized and the model is automized so the best values for Epoch and Batch optimization is find out and it is assigned. This leads to an increase in accuracy and reduced the loss and MSR (Mean Square Error). The Vanishing Gradient Problem is overcome by using ReLU (Rectified Linear unit) activation function. The formulae for ReLU is

R= {0, z}

{1 for z > 0}

{0 for z < 0}

The input of each neuron is passed to the activation function and there it is processed. ReLU overcoming the problem of vanishing gradient by keeping the derivative value as positive. So always there is a difference between Wold and Wnew in ReLU. Feedback connections were introducing LSTM and the best score for grid search is found out. By using standard Grid Search Value of Epoch and Batch size is automated and it is found out that for this dataset the best result was obtained for Batch size =20 and Epoch=20. Forward propagation is improved by adding the features of LSTM [12]. So a new memory part was introduced to store [12] the activation details of the hidden layers. Various steps involved in this model are described below:

- Choose the data set
- Convert all the input values to Vectorised format for standardization Find out the training values
- Initialise the model using Keras Classifier
- Three sub models are introduced, and, in each model, Optimal parameters are finding out using hyper parameter tuning in Grid Search Model.
- In model1 the best values for Epoch and Batch size was find out using Grid search in Keras Classifier. The model is tested with three batch sizes (10,20 and 30) and three epochs (10,20 and 30). All combinations were tested and the best accuracy will be find out for the combinations of batch size 20 and epoch 20.
- In Model2 Forward propagation is improved by introducing LSTM and Adam Optimizer. By automating the model best values for epoch and batch size was predicted. Learning rate and dropout rate is find out and best classifying accuracy is calculated.
- In Model3 ReLU activation function is introduced along with Softmax function. Best predicted result was observed.
- Stacked the three Models using tensor flow method and the accuracy matrix was printed. Logistic Regression is used for stacking the Models.

FINDING OPTIMAL PARAMETERS USING HYPER PARAMETER TUNING IN GRID SEARCH MODEL.

- In model1 the best values for Epoch and Batch size was find out using Grid search in Keras Classifier. The model is tested with three batch sizes (10,20 and 30) and three epochs (10,20 and 30). All combinations were tested and the best accuracy will be findout for the combinations of batch size 20 and epoch 20.
- In Model2 Forward propagation is improved by introducing LSTM and Adam Optimizer. By automating the model best values for epoch and batch size was predicted. Learning rate and dropout rate is find out and best classifying accuracy is calculated.
- In Model3 ReLU activation function is introduced along with Softmax function. Best predicted result was observed.
- Stacked the three Models using tensor flow method and the accuracy matrix was printed. Logistic Regression is used for stacking the Models.

RESULT AND DISCUSSION

From the observations it is found out that the accuracy is incremented in the new model comparing with the base classifiers. Here three base classifiers are used for the accuracy comparison. Naive Bayes, Logistic Regression, Multilayer Perceptron. From the observations it is found out that the automated hyperparameter optimization with stacked ensembled model increases the accuracy level.

Algorithm	Accuracy	Precision	Recall
Navie Bayes	76.3	75.9	76.3
Logistic	77.2	76.7	77.2

Regression			
MLP	78.3	75.0	74.0
Stacked Model	90.7	94.0	82.0

Table 1: Performance Comparison of Classifiers

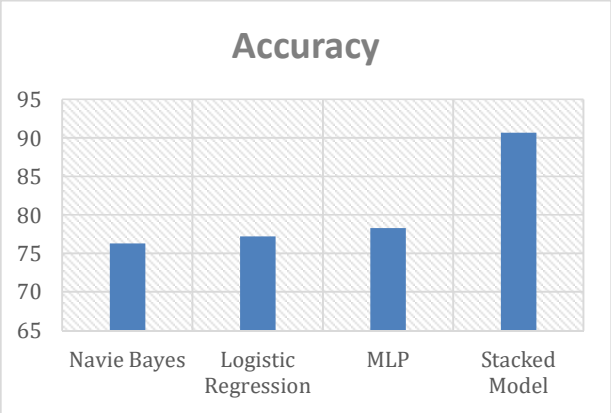


Figure 2: Classifier Performance based on Accuracy

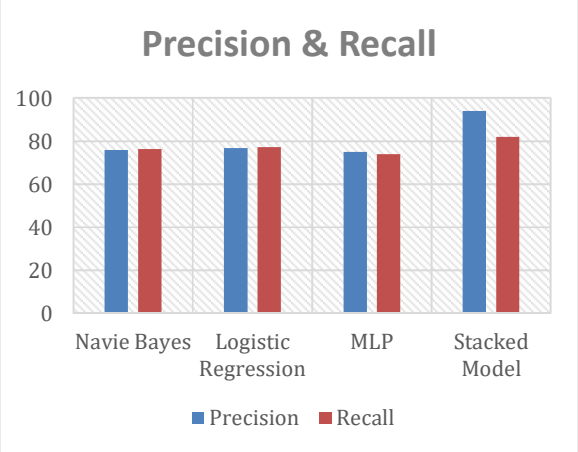


Figure 3: Classifier Performance based on Precision and Recall.

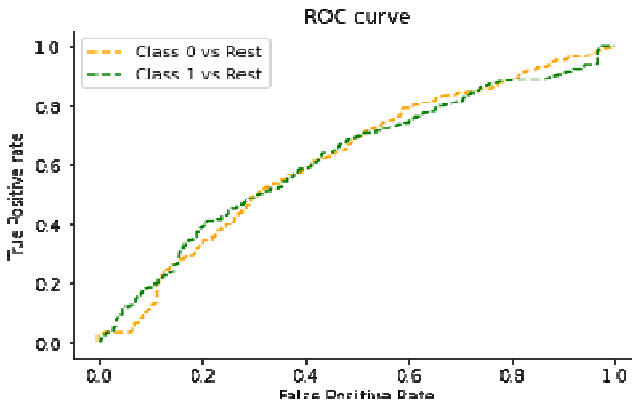


Figure 4: ROC Curve

ROC curve is a graphical plot that shows the system's diagnostic potential as its discriminating threshold is varied. The ROC curve is generated by plotting the true positive rate at different threshold settings against the false positive rate.

CONCLUSION

In this article we are introducing an ensembled hyper parameter tuning mechanism to tackle the deficiencies and improving the accuracy. For this purpose, we have used an automated stacked ensemble method which combines various hyper parameters. Grid Search Optimisation method is used, and three different models were created as base learners using Neural Network by combining various activation functions. The optimum value for each parameter is calculated and stored into an external file by each model. Three output files are created, and these output files are inputted to a logistic regression model which is a learning model. We have used LR Model as the learning model. It is found out that this model improves the accuracy to good extend.

REFERENCES

1.Roshan Birjais, Ashish Kumar Mourya,Ritu Chauhan, Harleen Kaur " Prediction and diagnosis of future diabetes risk : a machine learning approach". Springer.

2.Md.Maniruzzaman,JahanurRahman, BenojirAhammed, Md.Menhazul bedin "Classification and prediction of diabetes disease using machine learning paradigm", Springer

3. Atik Mahabub "A robust voting approach for diabetes prediction using traditional machine learning techniques", Springer.

4.www.geeksforgeeks.org/ml-chi-square-test-for-feature-selection

5.www.niddk.nih.gov/health-information/diabetes/overview/what-is-diabetes Retrieved.

6.www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death, Accessed 27th Jul 2018.

7.www.who.int/news-room/fact-sheets/detail/diabetes retrieved 27/07/2018.

8.www.diabetesdaily.com/learn-about-diabetes/what-is-diabetes/how-many-people-have-diabetes.

9.Changsheng Zhu, Christian Uwa Idemudia "Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques, ScienceDirect.
10.https://towardsdatascience.com/ways-to-detect-and-remove-the-outliers-404d16608dba
11.https://mc.ai/deep-learning-in-the-real-world-how-to-deal-with-non-differentiable-loss-functions/
12.Merijn Beeskma, Suzan Verberne, Antal van den Bosch Predicting life expectancy with a long short-term memory recurrent neural network using electronic medical records.
13.Sampath Kumar Palaniswamy, Hyperparameters tuning of ensemble model for software effort estimation, Springer.
14.Aramesh Rezaeian, Marzieh Rezaeian "Prediction of mortality of premature neonates using neural network and logistic regression", Journal of Ambient Intelligence and Humanized Computing, Springer.
15. AH. Habbi and M. Zelmat "Fuzzy Logic Based Gradient Descent Method with Application to a P1-type Fuzzy Controller Tuning: New Results", IEEE 2017.
16. Huaping Zhou, Raushan Myrzashova and Rui Zheng "Diabetes prediction model based on an enhanced deep neural network" EURASIP Journal on Wireless Communications and Networking, Springer.
17. Kalaiselvi, G. M. Nasira, "A New Approach for Diagnosis of Diabetes and Prediction of Cancer Using ANFIS", IEEE
18. Pahulpreet Singh Kohli, Shriya Arora "Application of Machine Learning in Disease Prediction", IEEE.
19. Abdulhakim Salum Hassan, I. Malaserene, A. Anny Leema , "Diabetes Mellitus Prediction using Classification Techniques", International Journal of Innovative Technology and Exploring Engineering (IJITEE)
20. Sai Poojitha Nimmagadda, Sagar Yeruva, Rakesh Siempu, "Improved Diabetes Prediction Model for Predicting Type-II Diabetes", International Journal of Innovative Technology and Exploring Engineering (IJITEE).

Certificate of Presentation

PSN COLLEGE OF ENGINEERING & TECHNOLOGY

(An Autonomous Institution)

Melathediyoor, Tirunelveli - 627 152.

International Virtual Conference on Computational Engineering (IVCCE- 2022)

Certificate of Appreciation

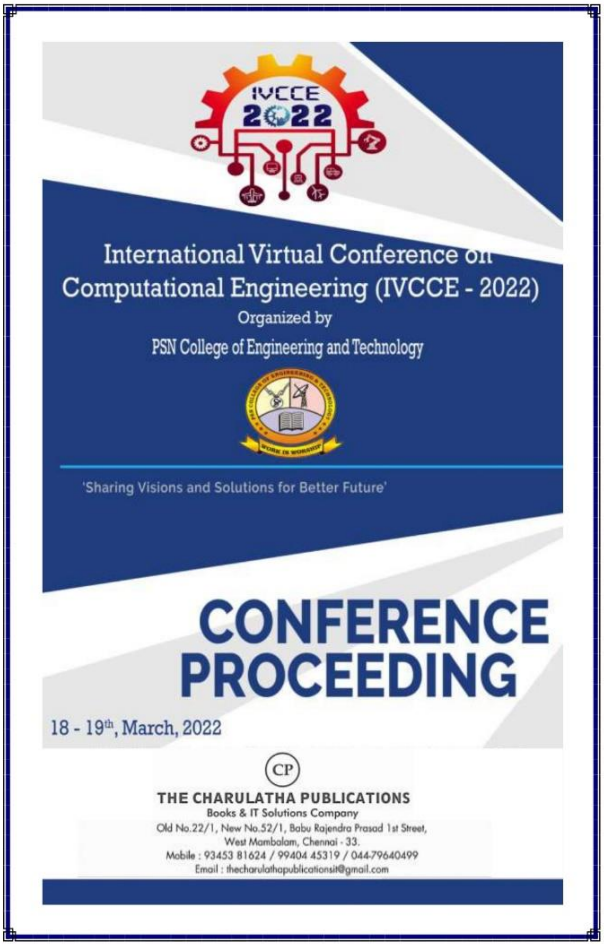
This is to certify that Ambily Merlin Kuruvilla of Saintgits College of Applied Sciences, Kottayam, Kerala,India has presented a paper entitled on Big Data in Cloud Computing Environment in the International Virtual Conference on Computational Engineering (IVCCE - 2022) on 18/03/2022 to 19/03/2022.

Convener
Dr.M.Vignesh kumar
Associate Professor

Executive Director
Dr.P.Selva kumar
Professor

Principal
Dr.V.Manikandan
Professor

Proceedings



IVCCE 2022

International Virtual Conference on Computational Engineering (IVCCE - 2022)

Organized by

PSN College of Engineering and Technology

'Sharing Visions and Solutions for Better Future'

CONFERENCE PROCEEDING

18 - 19th, March, 2022

CP

THE CHARULATHA PUBLICATIONS

Books & IT Solutions Company

Old No.22/1, New No.52/1, Babu Rajendra Prasad 1st Street,
West Mambalam, Chennai - 33.

Mobile : 93453 81624 / 99404 45319 / 044-79640499

Email : thecharulathapublicationsit@gmail.com

INTERNATIONAL JOURNAL OF
RECENT TRENDS IN MULTIDISCIPLINARY RESEARCH
May- June 2022, Vol 2(03), 01-05.
ISSN No: 2583-0368

Big Data in Cloud Computing Environment

Ansaba R B [1], Ambily Merlin Kuruvilla [2]
[1] *BCA student, Saintgits College of Applied Sciences, Pathamuttom, Kottayam, Kerala, India*
[2] *Assistant Professor & HOD, Department of Computer Applications, Saintgits College of Applied Sciences, Kottayam, Kerala, India*

Article Type: Research

OPEN ACCESS
Article Citation:
Ansaba R B[1],Ambily Merlin Kuruvilla [2], "Big Data in Cloud Computing Environment", International Journal of Recent Trends In Multidisciplinary Research, March-April 2022, Vol 2(03), 01-05.

Accepted date: May 12, 2022

Published date : May 14, 2022

Published by 5th Dimension Research Publication.

Abstract
In today's world, Big Data is an important area that is used in decision making and it processes huge volumes of data to address some query or pattern. Data is analysed through a set of algorithms, which differs depending upon the type of data, business's aim behind the analysis, and also other factors. But bigdata possess many challenges in terms of storing and processing data. Hence cloud computing which is another emerging technology is integrated with big data which provides better infrastructure for processing, storage for enormous data, and networking services.

Key Words: Big Data; Cloud Computing; Hadoop; Hdfs; Map Reduce

1. Introduction
Single Cloud computing is a powerful model and infrastructure that is distributed across the internet which process, manage and store the data. Cloud computing offers services for enterprise applications which centralizes both data storage and perform huge scale complex computing. It can reduce maintenance cost, provide less infrastructure and accelerate automation. [1]
Cloud services enables big data to analyse, manage and process the stored data in a more efficient manner. Through virtualization process integration of big data with cloud is the being achieved. Virtualization denotes the usage and sharing of resources independent of underlying hardware. Microsoft's Cloud Hadoop includes Azure Marketplace which comprise MapR and Azure Data Lake, which comprise Data Lake Store, Azure HDInsight, Data Lake Analytics as Azure cloud services. AWS includes versions of Hadoop, Spark, and Presto which operate on the data stored in Amazon Glacier and S3. Google's managed Hadoop include Cloud Dataproc and Spark cluster which uses GCP cloud services such as Big Query and Bigtable. [2]
Cloud platform provide rich productivity suites for database, data warehouse, collaboration, business intelligence, OLAP, and development tools. Big Data processing has many challenges relating with Data collection, analysis, sharing, research and visualization. Each of these processes need different techniques, infrastructure, and highly skilled professionals. Also, it cannot be done easily with traditional programs because of resource restrictions such as computing power and time, hence we need advanced algorithms and vast databases. And all these difficulties and barriers are much reduced as a result of integrating Big Data within cloud environment. [3]
Big data represents huge amounts of complex data which can be either unstructured or structured generated by multiple sources. The traditional relational databases are not sufficient to process and analyse data from multiple-sources, such as managing data related with record of transactions, customer behaviour, mobile phone and GPS navigation, etc. So, to deal with these kinds of complex data, cloud is employed, which serve as the storehouse where the processed outcome/data will be stored. Cloud computing approach is efficient because of having advanced technologies to handle the vast amount of data. This paper discusses an overall view of cloud computing and big data, their features, Relation and integration of big data & cloud, some **big data management tools in cloud**.

2. Cloud Computing
Cloud computing is a type of service-oriented computing where software and hardware are delivered as a service over the internet. Cloud is a combination of distributed and centralized system which includes virtualized servers, operating systems, applications, etc that are dynamically supplied. It provides services relating to storage, processing and sharing of data through visualized resources over the networks. Cloud platform is completely virtual to its users and require less effort from user to operate and manage its services. Important features associated with cloud includes scalability, on-demand delivery of resources, easy accessibility, cost-effective, flexibility and reliability. [4]

It has another major feature, Pay-as-you-use which means that users have to pay only for what they need at any given time. Advantages of cloud computing include:

- Data security
- Virtualized resources
- Easy and agile development
- Less maintenance cost
- Scalable data storage.
- Services in the pay-per-use model.

Cloud Service Models:

Service-oriented architecture of cloud supports "everything as a service" and hence offers their services as different models which are: [5]

Fig 1: cloud service models

- ***Platform as a Service (PaaS)****:*

In this service model, platform level elements such as project management environments, scalable and elastic runtime environment are provided. User can configure and install required software on the cloud. In short, PaaS provides the framework needed to build, deploy, test and manage software resources.

- ***Software as a Service (SaaS):***

It is a software distribution model where cloud consumers on internet can retrieve software applications and databases that are hosted by cloud service provider. If user does not have specific software or associated compatible hardware not installed on local computer, he/she can access directly from cloud. [6]

- ***Infrastructure as a Service (IaaS):***

Computing resources are provided to consumers by the IaaS in the form of infrastructure like virtual machines, servers, operating systems, network, hardware resources and storage on demand across internet . It provides completely virtualized computing infrastructure and provides an environment to deploy and run infrastructure including hardware and software in cloud environment.

Types Of Cloud:

Before transferring a business system into cloud, there is a need to consider many factors. There are four contrasting types of Clouds and three among them are basic types and hence mostly used. [7]

- ***Public Cloud:***

This type of cloud is available to the general public. General uses of public clouds include file-sharing, online office applications, application development and testing and web-based email. Public cloud infrastructure services are provided over the internet and hence open for everyone. Through public cloud, customers and users can easily access systems and shared

resources with low cost and high efficiency. Some examples of public cloud are Microsoft Azure, Amazon Elastic Compute Cloud(EC2), Google Cloud , Alibaba Cloud, Oracle Cloud Fast Connect .

- ***Private Cloud:***

It is also called "internal cloud" or personal data center computing.

It is deployed on a private network and are meant for the unique use of a particular company. This model provides highest level of security and data privacy as it permits only authorized users. They are more expensive than public cloud. Through this model, it is only able to access systems and services within an institution or an organization.

- ***Hybrid Clouds:***

It combines and integrates both private cloud and public cloud. Hence it allows cost-effective way for businesses to increase compute capacity on demand and better flexibility in terms of data transfer. Users or customers can develop and deploy applications using public cloud and at the same time offers higher degree of security through private cloud rather than using only a public cloud.

3. Big data

Big data refers to large amounts of data or compound datasets produced by various sources like sensors, mobile devices, social media and from three primary sources: machine data, social data and transactional data, in a very short duration of time. Such data are too large, fast growing and are difficult or impossible to process using traditional methods or conventional tools and techniques. Through deep analysis and efficient processing by various Data Analytics methods, valuable information can be extracted from big data. [8]

Characteristics: Big data is characterized mainly by five Vs which are: Volume, Veracity, Variety, Value and Velocity.

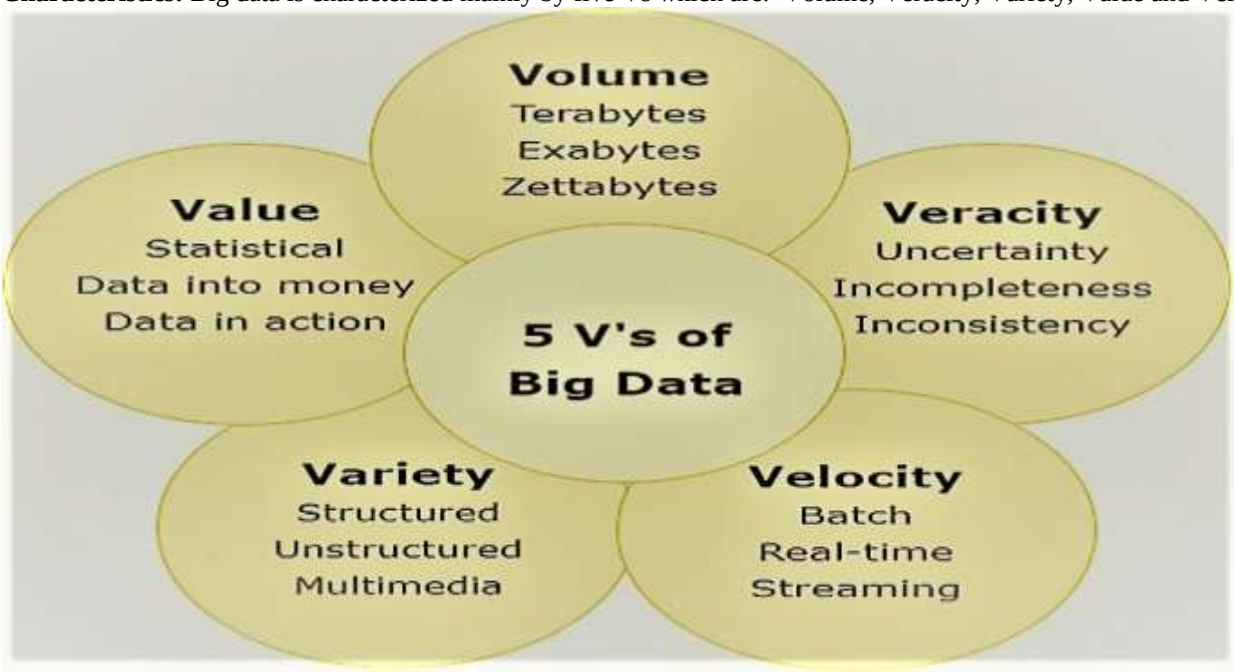


Fig 2: characteristics of big data

Volume:

It denotes incredible amount of data which are generated and stored in Gigabytes (GB), Zettabytes (ZB), and Yottabytes (YB). In coming years, the volume will rise significantly as data is being created every second from various sources like social media platforms, smart (IoT) devices, networks, machines and so on. [9]

Variety:

It refers to various kinds of data gathered from different sources. Data generated can be of different formats which can be structured, unstructured, semi structured or a mix of all these three. It can include different forms of data such as financials, logs files, social media updates, images, videos, text messages, audio, etc.

Veracity:

It denotes "compliance with truth or fact" and refers to overall quality and reliability of the data source. Low veracity can negatively affect the accuracy of the results.

Value:

It denotes the final value obtained after processing of data and produced during analysis which helps in decision making. To obtain value firstly mine data which refers to the process of conversion of raw data into useful data. Next on this retrieved data, analysis is done.

Velocity:

It denotes the speed at which data is being created, generated, collected and analysed. Velocity also associated to how fast big data is going to be processed.

Examples of data generated with high velocity include Facebook posts, data from sensors and mobile devices.

Big Data Advantages:

- Real-time monitoring of product price optimization, business and market
- Greater innovations and lifesaving applications in the healthcare industry and public health with availability of record of patients.
- Real time communication regarding customer requests, their queries and problems.
- Helps in quicker and better decision making

Challenges of Big Data:

- Difficult to Manage large volumes of data as there is always a lot of raw data to store and analyse.
- Lack of workers with adequate big data skills and talent.
- There is a chance to make wrong decisions due to unevenness of data quality and it is difficult to determine which source of data is correct.
- Poor data scalability, reliability and runtime quality issues.[10]

4. Integration of Bigdata in Cloud

Cloud platform provides one of the best environments for efficient bigdata processing and real time analysis in a cost-effective way. It has greatly improved Big Data analysis, resulting in better findings and hence decision making. Cloud environment provide services to analyse and process bigdata by breaking huge volume of information into smaller units and each of them can be processed independently in different servers. Through remote multi servers and dynamic parallel resource allocation, it is possible to handle massive amount of data accordingly in cloud environment. Integration with cloud make big data resources more monitored, productive, compliant and simpler. [11]

Cloud providers like Google Cloud Platform, Amazon Web Services, Microsoft Azure, IBM, Oracle, Salesforce, etc. provide important factor: scalability which is required for bigdata handling and processing.Another important factor is the data security and privacy which cloud platform offers. It provides more scalable and elastic Private Cloud Solution thereby a safe environment to keep big data and its computation. To store data on the cloud, a key is given to all its users and data can be accessed only by using that key.Cloud error localization is a technique which is used to identify and monitor error in big data storage and also handles bad performance of server. [12]

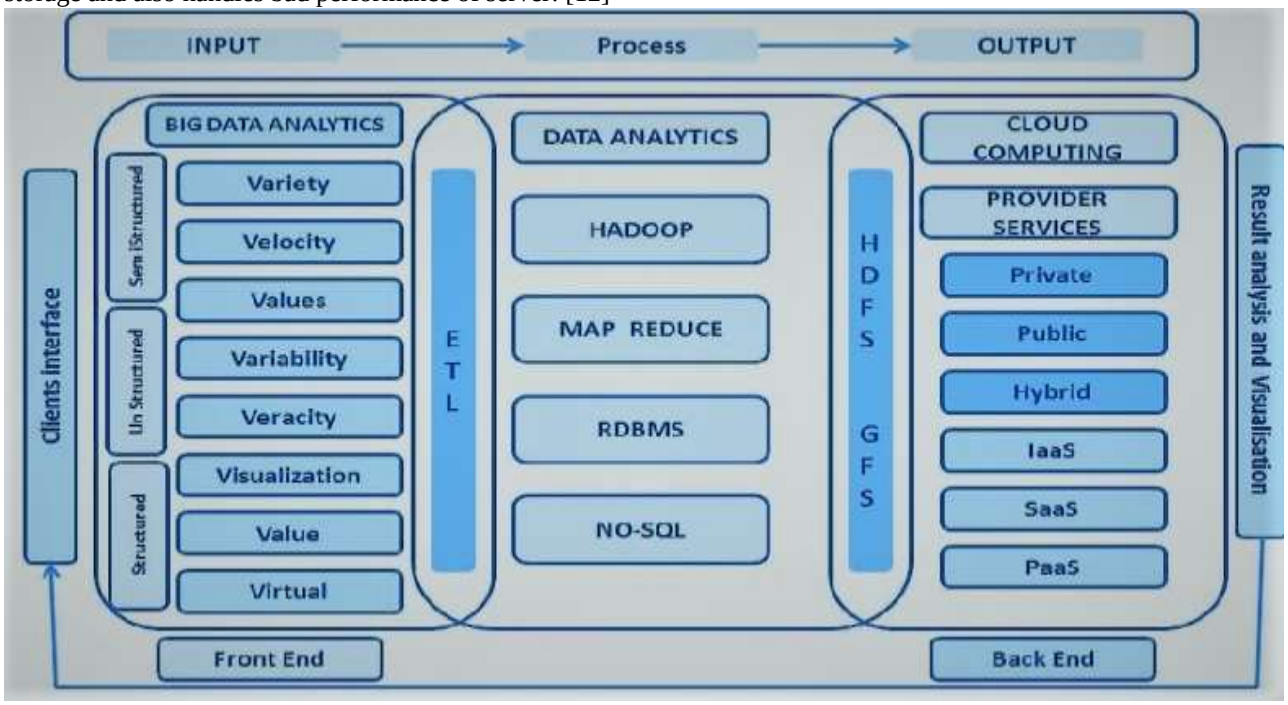


Fig 3: Relation between cloud and big data

Big data management tools in cloud:

• **Hadoop**:

Hadoop is a part of Apache project and it is a freely available java-based programming framework. Hadoop enables processing of large sets of data on a cluster of servers and applications consisting terabytes of data. So, even if some node fails, Hadoop supports with rapid transfer rates. Hadoop consists of higher-level declarative languages for big data analysis pipelines and query writing. Hadoop mainly composed of HDFS and MapReduce.

• **HDFS**

HDFS is a file system used to store or span all the nodes in a Hadoop cluster for data storage. Thereby it improves reliability and support security. HDFS usually splits files into blocks which in turn is stored on the server. Thereby it maintains reliability by duplicating data across multiple hosts combining parallel processing technique. [13]

- **MapReduce**

This is a framework which helps in writing applications that process and generates large datasets on a cluster with parallel or distributed algorithm. At first, breaking Big Data into small subunits takes place which in turn are analysed and processed by Map jobs in parallel. Map () method consists of acquiring, filtering & categorizing datasets. Reduce () method consists of final result generation and locating associated summaries. [14]

- **NoSQL**

NoSQL (Not Only SQL) systems provides systematic way to store and replicate data, giving out retrieval and appending operations from the data. These databases are not bound by the confines of a fixed schema model instead each are deployed as a cluster of nodes. Examples of NoSQL systems include Amazon DynamoDB, Azure Cosmos DB, MongoDB, Cassandra, CouchDB, and HBase.

5. Conclusion

This paper presented how cloud computing helps in analysing, storing and processing big data. Big data and cloud together comprise an integrated model of distributed network technology.Cloud supports big data in terms of security of data, encryption, data integrity, data transformation, data heterogeneity, data quality and others.
Even though there are challenges regarding integration with cloud such as scalability, availability and problems with bandwidth for data transfer, Solutions are constantly being developed by cloud providers for the efficient use of big data on cloud. So, the integration and application of big data in cloud will have a huge impact and continue to grow in the following years.

References

[1] Beri, R. & Behal, V. (2015). Cloud Computing: A Survey on Cloud Computing. Int. J. Comput. Appl., Vol. 111, pp. 19–22.
[2] Sheetal Singh, Vipin Kumar Rathi, Bhawna Chaudhary, "Big Data and Cloud Computing: Challeges and Opportunities",International Journal of Innovations in Engineering and technology, Vol. 5(4), August 2015.
[3] Gupta, H. & Mohania, M. (2012). Cloud computing and big data analytics: What is new from databases perspective? in Big Data Analytics. BDA 2012. Lecture Notes in Computer Science, Vol. 7678, pp. 42–61 (Springer Berlin Heidelberg).
[4] Fonseca, N., & Boutaba, R. (2015). Cloud services, networking, and management. John Wiley & Sons.
[5] K. Kaur, "A Review of Cloud Computing Service Models", International Journal of Computer Applications, Vol.140, No.7, pp.15-18, 2016.
[6] J. Srinivas, K.Venkata Subba Reddy and Dr. A. Moiz Qyser, "Cloud Computing Basics", International Journal of Advanced Research in Computer and Communication Engineering, Vol.1(5), 2012.
[7] Venters, W., Whitley, E.A.: A Critical Review of Cloud Computing: Researching Desires and Realities. J. Inf. Technol. 27, 179–197 (2012).
[8] D.P. Acharjya, Kauser Ahmed P, " A Survey on Big Data Analytics: Challenges, Open Research Issue and Tools", International Journal of Advanced Computer Science and Applications, Vol. 7(2), 2016
[9] N. Elgendy and A. Elragal, "Big Data Analytics: A Literature Review Paper," in Advances in Data Mining.Applications and Theoretical Aspects, 2014, pp. 214–227.
[10] A, Katal, Wazid M, and Goudar R.H. "Big data: Issues, challenges, tools and Good practices.". Noida: 2013, pp. 404 – 409, 8-10 Aug. 2013.
[11] Bautista Villalpando, L. E.; April, A. & Abran, A. (2014). Performance analysis model for big data applications in cloud computing, Vol. 3, pp. 1–20
[12] Agrawal, Divyakant & Das, Sudipto & Abbadi, Amr. (2011). Big Data and Cloud Computing: Current State and Future Opportunities. ACM International Conference Proceeding Series. 530-533. 10.1145/1951365.1951432
[13] K, Chitharanjan, and Kala Karun A. "A review on hadoop — HDFS infrastructure extensions.". JeJu Island: 2013, pp. 132-137, 11-12 Apr. 2013.
[14] L. Zhao, Z. Zhou," Cloud computing model for big data processing and performance optimization of multimedia communication", Computer Communications (2020)