Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 07, July-2023

Efficient prediction of internet service provider using machine learning

Sanju Patil
School of ECE
KLE Technological University
Hubli, INDIA

Suneeta V. Budihal
School of ECE
KLE Technological University
Hubli, INDIA

Saroja V. Siddamal
School of ECE
KLE Technological University
Hubli, INDIA

***Abstract*—In the new trend of process of improving the network services there are lot of challenges faced in communication domain among which selection of best internet network is also a major one. To overcome this we designed a machine learning model which predicts the best service provider among given data and helps us to improve the service by providing a feedback analysis on the predicted information based on several characteristics. Machine learning algorithms are being used here as it can provide a broader class of several alternative analysis methods which are best suited to modern data set.**

I. Introduction

We are currently contributing to the definition of what 5G networks will resemble. Multiple work items that will result in the definition of a novel 5G radio and architecture have already been started by the 3GPP [1].Numerous white papers detailing the leading vendors' perspectives on 5G networks and architectures are being published. In order to fund research for 5G networks, the EU commission has established a significant 5G Infrastructure Public Private Partnership (5GPPP) programme [2]. It is evident from all of these convergent perspectives that network management in future 5G networks will have to deal with a completely new set of issues.

In this situation, it is already commonly acknowledged that new protocols must be set up in order for the network to become more intelligent, self-aware, and self-adaptive. Since Release 8, the Self-Organising Network (SON) has been a part of 4G LTE networks, which is the first step in this approach. However, given the enormous complexity of these networks, this idea needs to be refined further for 5G. As we previously noted in, control and management functions generate a significant amount of data during routine operations in 4G, and more data is gathered in 5G as a result of the densification process, heterogeneity in layers and technologies, the added complexity of control and management in NFV and SDN architectures, and the growing importance of M2M and IoT communications.

This paper proposes to predict the best internet service provider using machine learning algorithms. The paper involves a series of steps carried out, starting with understanding the problem definition and reviewing the data set, setting an end goal for a desired problem and carried out by listing the alternate solutions and selecting the best suited solution. Searching for the respective algorithms and understanding the working of algorithms to propose a solution for a our desired problem statement and then implementing the solution by testing and training the model and evaluating the accuracy and deploying the model.

Every machine learning project is carried out through these steps: Although the identified solution is suitable for the problem statement, but trying different models is required in order to know the variations in accuracy to build a proper model with best prediction analysis. This paper aims at improving the existing algorithms with some modifications so as to make the algorithms free of errors. Hence, it is essential for securing, stabilizing and increasing the efficiency. As virtualization is a great concern in this era, there is a huge demand for the good network speed in the society. This paper helps the service provider to implant the tower in required area as well as help the customer to choose the best network provider in their respective areas.

A. Literature survey

Authors in paper [3] have proposed the concept of mobile network tool based on prediction of data analytic, where they mainly focused on best network planning. It gives a proper definition to the service provided to the users and the resources used by the customers. This framework worked mainly in two steps. First, they proposed a model of service through the analysis of data collected from the networks in the form of different measures. Second, they changed the parameters and analysed the impact on QoS based on the previous learning. In this way, the performance is optimised to meet the targets focusing on Random Forest algorithms. Authors in paper[4] have given brief overview of Random Forest Algorithm. They have also discussed the key feature of Random Forest Algorithm i.e., node size, the total number of trees and features sampled, etc. In general they have highlighted the importance of Random Forest algorithm i.e., proof-based and vote-based.

Authors in paper [5] have given a brief overview on feature extraction method that is based on deep belief networks and random forest algorithm; This algorithm is based on multi layer neural network to minimise the dimension of the provided data, and then, Random forest algorithm is applied. The algorithmic model consists of the following:

- Data acquisition using wavelet denoizing
- Feature extraction using deep neural network method

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 07, July 2023

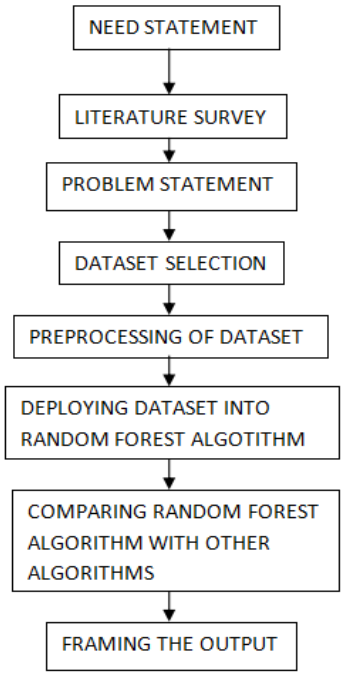

Fig. 1: Flow diagram for the proposed methodology

- Signal recognition using classification method.

After collection of data and preprocessing the data, DNN is used for extraction of feature vectors from training and testing dataset. Then RF model is used for classification of training dataset and use testing dataset for validation purpose. Authors in paper [6] have given the importance to the variables that are important in predicting bovine viral diarrhea virus. The method in random forest algorithm has properties which makes it appealing for classification problems.

Authors in paper [7] have proposed transmission rates and other physical characteristics of the network that are measured and improved and to some extent analysis of the network [8-10]. Traffic load is very important for optimisation of the proposed model and to control the services provided by the network. These services are important for control mechanism on network and also maintaining the maximum utilization of the service. Resource allocation [11-12] is optimised in such a way that it imitates with QoS .

B. Problem Statement

To Predict the best Network provider/ best Internet service provider across India in different states according to their network performance. This is a data set of government of India collected by TRAI using Myspeed application. The data is sampled from 1.3 million devices on which Network speeds were measured by Myspeed application of TRAI. The samples are taken from all the states of India, from various Service Provider. Characteristics present in the dataset are Service Provider, Technology Test type, Data speed(mbps), Signal strength and LSA. The dataset contains roughly equal no of download and upload tests, collected from March 2018 from various states of India and consisting of various service providers. Apart from the data sets for 4G and 3G network, the data also consists of signal strength while the speeds were measured.

C. Objectives

- Develop a Prediction based model.
- Implement Random Forest algorithms using Machine learning.
- Training and Testing of implemented model using machine learning.
- Deployment and performance check of Random Forest algorithm using different.

The above propose model helps in various different ways which acts an application for society to improve their services

- Customers can Easily choose the best service provider by analyzing the nature of predicted service provider the customers can choose the best network for their area.
- Improving the services by service provider by the prediction data available as a result of prediction the service providers can study the data and can improve the service for particular area or state as required.

II. Proposed System Framework

The Figure 2 represents the System design of proposed model.

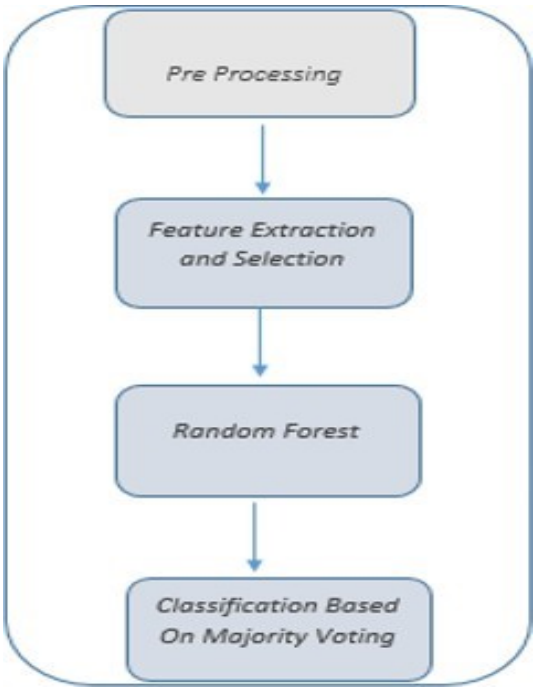

Fig. 2: Architecture of System Design

The given problem in this case as indicated will be solved by using Machine Learning model.specifically Random Forest implemented using python programming language with the help of multiple inbuilt libraries. Hence, as described in the

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 07, July-2023

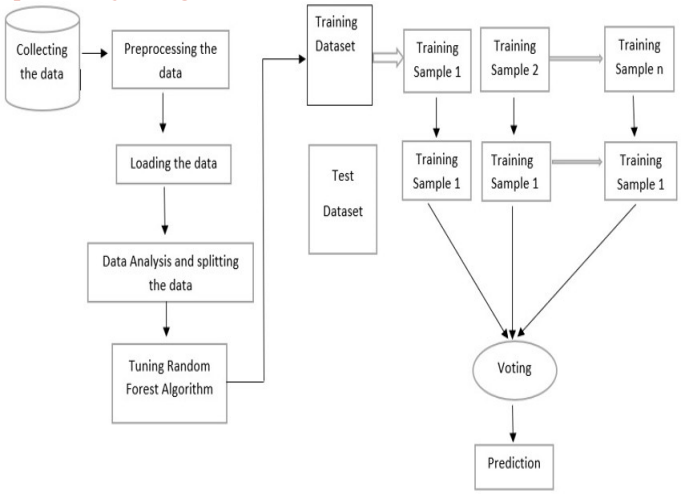

Fig. 3: Flowchart of Random Forest

following section we discuss the details of system design of architecture capable to solve the given problem and the same using python and libraries for the purpose.

A. Analysis of Proposed model

Supervised learning consists one of the algorithm named Random Forest Algorithm which is usually used for classification problems rather than regression problems. Random Forest algorithm is made up of multiple decision trees, as number of trees increases the more is the robust forest. Random Forest creates decision tree and predicts the output for each tree, then using the voting method it takes the best solution among all the trees. This method is considered because it is better to take output from multiple trees rather than selecting the solution from single decision tree. It increase the accuracy also overcomes over-fitting problem.

B. Flow Chart

Figure 3 shows the flow of Random Forest Algorithm. The data set is split into two segments training data (70%) and testing data (30%). Then N number of samples are taken from training data set to train the model. Then voting is done among all the decision trees created from those training samples. Then the decision tree with more numbers of votes is selected as the best prediction method.

C. Functional Block Diagram

Figure 4 is the functional block diagram of Random Forest The working of Random Forest Algorithm as sequence diagram of the proposed model is shown in Figure 4. Initially there is a group of decision trees created and voting is done among them by then the training and testing data is split then variables are chosen and stop conditions are applied for each chosen variable at next splits.then sorting of variables is done then index is calculated at each split and prediction error is calculated.

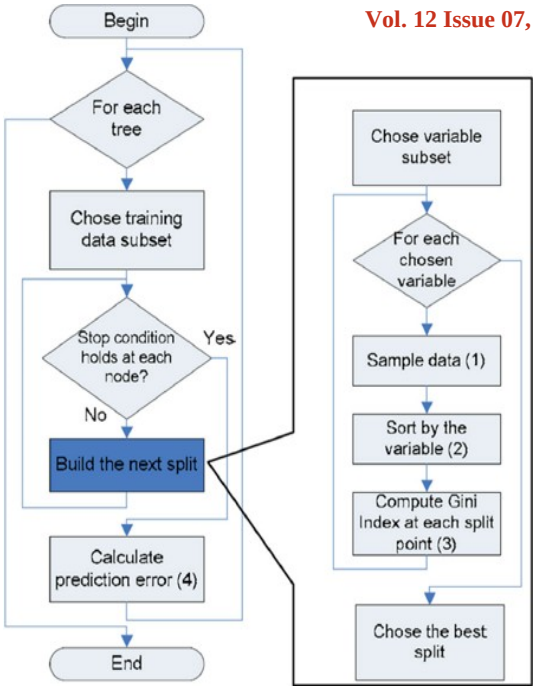

Fig. 4: Functional Block Diagram of Random Forest

- Considering the training samples to be N and testing data samples to be M
- m number of inputs is used to determine the decision at node of the tree where M should be greater than m
- Choosing training samples for the tree
- For each node of tree randomly m variables are chosen.
- Taking results from each tree and best solution is chosen by voting method.

For prediction a new sample is pushed down the tree. It is assigned the label of the training sample in the terminal node it ends up in. This procedure is iterated over all trees in the ensemble, and the average vote of all trees is reported as random forest prediction.

D. Advantages of Proposed Model

This model runs efficiently on large data set.

- One of the advantage is that it handle thousands of input variables without variable deletion.
- The model can provide estimates of what variables are important in the classification.
- This model offers an experimental method for variable detection.

III. Implementation Details

In this section the details of requirements of the system are given. The implemented paper is capable of predicting the best Internet service provider across India in different states according to their network performance.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 07, July-2023

A. System Requirements

The system should be able to implement Machine learning Algorithm

- The system should be able to implement Random Forest algorithms.
- The system should be able to implement the SVM algorithm
- The system should detect and Predict the result with maximum accuracy. The non-functional requirements are:
- The system should be able to display plots that are the outputs of running code cells
- The system should reduce the resource consumption by considerable amount.

B. Software Requirements

- Windows/Mac
- Jupyter Notebooks/Google colab
- Python 3.6 with Keras, Matplotlib, Numpy, Pandas, sklearn Libraries.

C. Hardware requirements

- Minimum 8GB RAM
- 512GB hard disk
- Minimum core i3 processor

D. Random Forest Algorithm

These steps represent the working of algorithm in creating the Random Forest model:
Step1: Begin
Step2: Collecting the data from TRAI
Step3: Preprocessing of data using different functions such as concat, drop, getdummies
Step3: Providing certain weightage for training and testing data
Step4: Training the dataset with random frorest algorithm
Step 5: Testing the results on testing dataset where accuracy is determined
Step 6: Comparing the Random forest algorithm with SVM and Neural networks to differentiate the accuracy of each algorithm with respect to random forest algorithm
Step 7: End

E. Optimization

Optimisation is a technique of finding set of inputs which leads to maximum accuracy and minimum function evaluation hence performing better. Many machine learning algorithms face problem from fitting the logistic regression to training ANN. The purpose of optimisation is to get the best design comparing to set of constraints or criteria. These include strength, longevity, productivity, reliability, utilization and efficiency. Optimisation is an important tool for making decisions and analysing the physical system. Optimisation is also defined as finding best solution from all the feasible solutions. Program or software optimisation is defined as process of modification of software system to work more efficiently with minimal resources.

No.	Algorithm	Training Accuracy	Testing accuracy
1	SVM	87.29%	80.5%
2	Random Forest	89.7%	87.5%

TABLE I: Comparison of algorithm accuracy

Optimisation technique used in this paper is hyperparameter tuning. Hyperparameters refers to the parameters which helps in defining the model architecture. Some of the hyperparameters are constraints, learning rate. Hence searching of ideal hyperparameters which play very important role in increasing the efficiency and accuracy of the proposed model.

IV. Results and Discussions

The following outputs were obtained when trained the model with Random Forest, SVM and Neural Networks Algorithms.

```
x_train (786431, 32)
x_test (262144, 32)
y_train (786431, 1)
y_test (262144, 1)

In [17]: from sklearn.ensemble import RandomForestClassifier
         clf = RandomForestClassifier(max_depth=15)
         clf.fit(x_train, y_train)
         y_pred_clf = clf.predict(x_test)
         clf.score(x_test, y_test)

/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:3: DataConversionWarning: A column-vector y was passed when a 1d a
rray was expected. Please change the shape of y to (n_samples,), for example using ravel().
  This is separate from the ipykernel package so we can avoid doing imports until

Out[17]: 0.8729782104492188

In [18]: y_pred_clf = y_pred_clf.reshape([262144,1])

In [19]: y_pred_clf = pd.DataFrame(y_pred_clf)

In [43]: parameters = [3, 5, 7, 9, 11, 13, 15, 17, 19, 35, 55]
         accuracy= list()

         for i in parameters:
           clf = RandomForestClassifier(max_depth=i)
           clf.fit(x_train, y_train)
           y_pred_clf = clf.predict(x_test)
           accuracy.append(clf.score(x_test, y_test))
           print(accuracy[-1])

         accuracy
```

Fig. 5: Random Forest Accuracy

	A	B	C	D	E	F	G
1	JIO	4G	download	24968	-88	Delhi	
2	JIO	4G	download	36353	-88	Delhi	
3	JIO	4G	upload	2437	-88	Delhi	
4	JIO	4G	upload	3160	-88	Delhi	
5	JIO	4G	upload	7766	-68	Kolkata	
6	JIO	4G	upload	7191	-68	Kolkata	
7	JIO	4G	upload	7656	-66	Kolkata	
8	JIO	4G	upload	7542	-66	Kolkata	
9	JIO	4G	upload	7185	-66	Kolkata	
10	AIRTEL	4G	download	0	-69	Andhra Pradesh	
11	IDEA	4G	download	7	-79	Maharashtra	
12	IDEA	4G	download	16007	-75	Maharashtra	
13	IDEA	4G	download	42602	-77	Maharashtra	
14	IDEA	4G	download	17626	-77	Maharashtra	
15	IDEA	4G	download	2811	-81	Maharashtra	
16	IDEA	4G	download	20358	-79	Maharashtra	
17	IDEA	4G	download	3108	-67	Maharashtra	
18	IDEA	4G	download	13251	-79	Maharashtra	
19	IDEA	4G	download	23240	-75	Maharashtra	
20	JIO	4G	download	53497	-64	UP East	
21	JIO	4G	upload	3889	-64	UP East	
22	JIO	4G	download	76807	-69	Rajasthan	
23	JIO	4G	upload	8682	-69	Rajasthan	
24	JIO	4G	download	22636	-70	UP East	

Fig. 6: Sample data set used for training model

The trained model has an accuracy of 87.5 % with a prediction method implemented with RF Algorithm. When

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 07, July-2023

tried with different model comparing SVM with Random Forest for results analyzes. We plotted a confusion matrix for SVM model and also got an accuracy of 80.5 %.

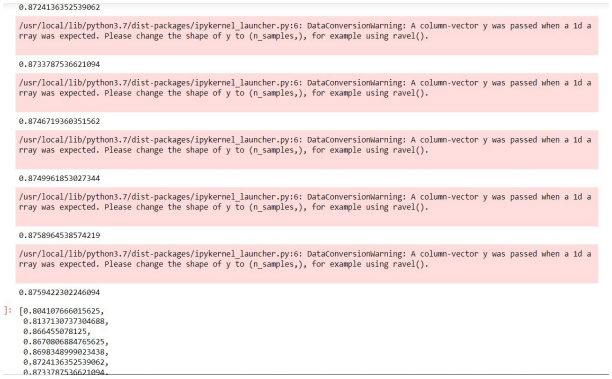

Fig. 7: Multiple Accuracy of the test cases for RF Model

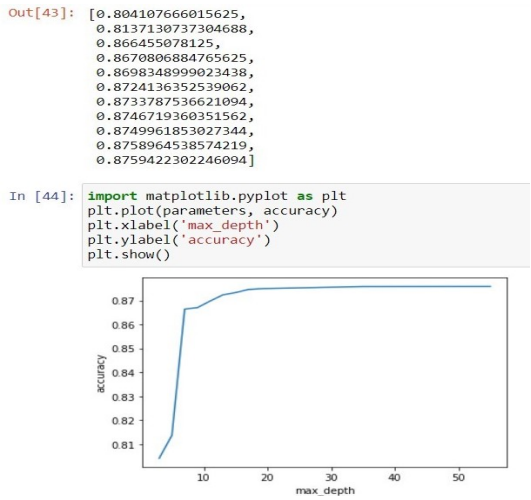

Fig. 8: Output plot over maximum depth and accuracy of RF Model

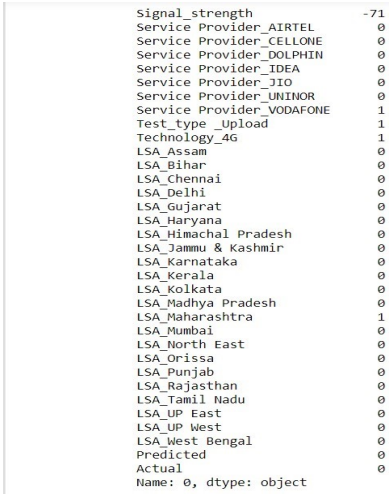

Fig. 9: Correct Prediction of test case for RF Model

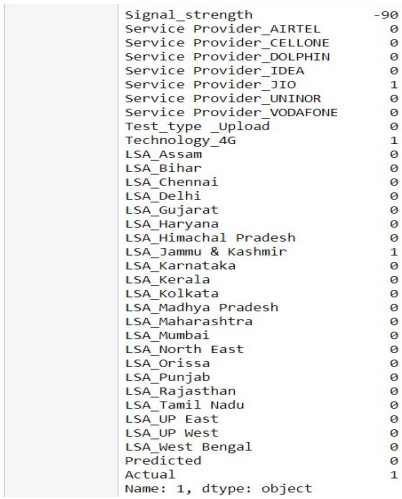

Fig. 10: False Prediction of test case for RF Model

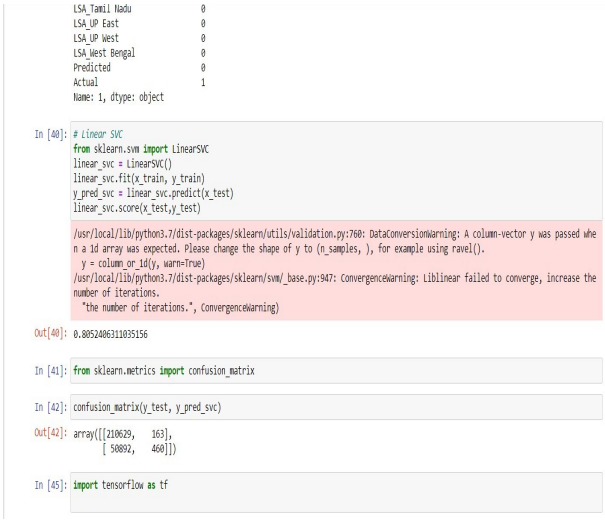

Fig. 11: Conclusion matrix for SVM Model

So Random Forest was having more accuracy then that of SVM model. When we tried comparing Random Forest with Neural Network we plotted graphs between epoch vs loss and epoch vs accuracy and the accuracy was less compared to that of Random Forest so we choose Random Forest as model of implementation. The Table 1 represents the comparison of accuracy of models.

V. Conclusion

In Machine Learning for an algorithm with sufficient resources, it is important to get a fair chance of prediction. So far in this paper we focused on how to use machine learning algorithms and tools, By developing a prediction based model on best network service provider. Initially the dataset was taken from TRIA and some preprocessing of the dataset was carried out then an application of random forest algorithm was implemented where random forest regressor algorithm was applied. Which leads an accuracy of 87.5 % others

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 07, July 2023

```
Epoch 3/10
24576/24576 [==============================] - 40s 2ms/step - loss: 0.3027 - accuracy: 0.8714 - val_loss: 0.2985 - val_accurac
y: 0.8708
Epoch 4/10
24576/24576 [==============================] - 38s 2ms/step - loss: 0.2993 - accuracy: 0.8724 - val_loss: 0.2972 - val_accurac
y: 0.8725
Epoch 5/10
24576/24576 [==============================] - 38s 2ms/step - loss: 0.2982 - accuracy: 0.8724 - val_loss: 0.2971 - val_accurac
y: 0.8724
Epoch 6/10
24576/24576 [==============================] - 38s 2ms/step - loss: 0.2979 - accuracy: 0.8721 - val_loss: 0.2966 - val_accurac
y: 0.8734
Epoch 7/10
24576/24576 [==============================] - 38s 2ms/step - loss: 0.2965 - accuracy: 0.8732 - val_loss: 0.2937 - val_accurac
y: 0.8737
Epoch 8/10
24576/24576 [==============================] - 38s 2ms/step - loss: 0.2960 - accuracy: 0.8726 - val_loss: 0.2922 - val_accurac
y: 0.8741
Epoch 9/10
24576/24576 [==============================] - 38s 2ms/step - loss: 0.2939 - accuracy: 0.8736 - val_loss: 0.3051 - val_accurac
y: 0.8709
Epoch 10/10
24576/24576 [==============================] - 39s 2ms/step - loss: 0.2956 - accuracy: 0.8729 - val_loss: 0.2949 - val_accurac
y: 0.8729
```

Fig. 12: Accuracy for SVM Model

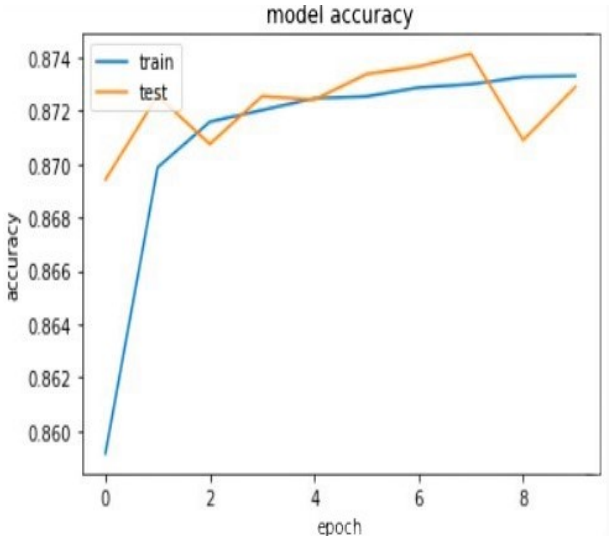

Fig. 13: Plot of Model Accuracy vs Epoch for CNN Model

algorithms like SVM and Neural Networks were also tried for a defined problem statement. Machine learning is one of the leading carrier choices at present where projects like these can be implemented to create a well and easy environment for Telecommunication industries to obtain the prediction based data of best and worst network service provider in order to improve their services an provide a better service.

REFERENCES

[1] 3GPP on track to 5G, http://www.3gpp.org/ news-events/3gpp-news/1787-ontrack 5g.
[2] 5GPPP.The 5G Infrastructure Public Private Partnership, https:// 5g-ppp.eu/.
[3] Lorenza Giupponi and Josep Mangues-Bafalluy, “A Mobile Network Planning Tool Based on Data Analytics,” Volume 2017 https://www.hindawi.com/journals/misy/2017/6740585/
[4] Ke Li, Nan Yu, Pengfei Li, Shimin Song, Yalei Wu, Yang Li, Meng Liu, “Multi-label spacecraft electrical signal classification method based on DBN and random forest,”May 2017 https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0176614
[5] Serkan Balli, Ensar Arif segbas, Musa Perkol, Humanactivity recognition from smart watch sensor data using a hybrid of principal component analysis and random forest algorithm” November 2018 https://journals.sagepub.com/doi/10.1177/0020294018813692
[6] Chuanting Zhang, Dongfeng Yuan, ”Fast Fine-Grained Air Quality Index Level Prediction Using Random Forest Algorithm on Cluster Computing of Spark” https://www.researchgate.net/publication/281061339
[7] Sajib Kabiraj, M. Raihan, Nasif Alv, Marina Afrin, Laboni Akter, Shawmi Akhter Sohagi, Etu Podder, ”Breast Cancer Risk Prediction using XGBoost and Random Forest Algorithm” 2020 International Conference, https://ieeexplore.ieee.org/document/9225451
[8] Suneeta, V.B., Purushottam, P., Prashantkumar, K., Sachin, S., Supreet, M. (2020). Facial Expression Recognition Using Supervised Learning. In: Smys, S., Tavares, J., Balas, V., Iliyasu, A. (eds) Computational Vision and Bio-Inspired Computing. ICCVBIC 2019. Advances in Intelligent Systems and Computing, vol 1108. Springer.
[9] Dai Chunni, ”SVM Visual Classification Based on Weighted Feature of Genetic Algorithm” 2015 sixth International Conference, https://ieeexplore.ieee.org/document/7462735
[10] Maniyar, H.M., Budihal, S.V. (2020). Plant Disease Detection: An Augmented Approach Using CNN and Generative Adversarial Network (GAN). In: Badica, C., Liatsis, P., Kharb, L., Chahal, D. (eds) Information, Communication and Computing Technology. ICICCT 2020. Communications in Computer and Information Science, vol 1170. Springer, Singapore.
[11] Pavaskar, S., Budihal, S. (2019). Real-Time Vehicle-Type Categorization and Character Extraction from the License Plates. In: Mallick, P., Balas, V., Bhoi, A., Zobaa, A. (eds) Cognitive Informatics and Soft Computing. Advances in Intelligent Systems and Computing, vol 768. Springer, Singapore.
[12] Suneeta V. Budihal, Rajeshwari M. Banakar, Evidence-based dynamic radio resource allocation to mitigate inter cell interference employing cooperative communication, IET Communications, Volume 14, Issue 12, July 2020, Pages 1848-1857