OPEN ACCESS

International
Journal
Of Modern Engineering Research (IJMER)

Flight Delay Prediction and Error Analysis Using Machine Learning

A.VEERAMANI [1], PATHAN IRUFHAN KHAN[2]

#1 M.Tech Scholar (CSE), Department Of Artificial Intelligence & Data Science,
#Associate Professor, Department of Artificial Intelligence and Machine learning in Kakinada Institute of Engineering and Technology-II, AP, India.

***Abstract** - Air transportation, scheduled to perform at the time when the flight has an important place in transportation systems, it is necessary to ensure the passengers' comfort and controllability of the operating costs. The Weather of flight delays experienced in air traffic density, accident or closed runway, the conditions that could lead to increasing the distance between planes and to live in ground services such as delays. In this study, data recovery from various sensors at the airport and from that estimate delays in flights with the flight information using an artificial neural network model has been targeted for improvement.*

***Key Words:** RBFN - Radial basis neural network, BPN - Backpropagation neural network, Binary Classification, MLP*
- Multilayer Perceptron

I. INTRODUCTION

The study of this paper mainly focuses on predicting flight delays based on historical data. The accuracy of the model during & after the process has to be maximum for better prediction. The paper explains the implementation of models that have been trained and tested by the historical data and can predict flight delays. We will be doing a comparative study between two algorithms after gaining the necessary references from this paper.

1.1 Problem Statement

Flight delays can be very annoying to airlines, airports, and passengers. Moreover, the development of accurate prediction models for flight delays became very difficult due to the complexity of air transportation flight data. In this project, we try to resolve this problem with approaches used to build flight delay prediction models using BPN and Radial Basis Function.

1.2 Objectives

- The objective of the proposed system is to predict the delays of flights. The system will be using 30% of the data for training the classification models and on the basis of that further delays can be predicted.

- This prediction will be done using Backpropagation network and Radial Basis function and in the end, the one with most accuracy will be considered as an efficient model and employed.

1.3 Scope

- This project has a large scope as it has many features which help in understand and modify it.
- The project can be further extended to predict flight cancellations as well it can be used for prediction over a larger period of time by training the models over a stipulated period.

II. MODEL REPRESENTATION

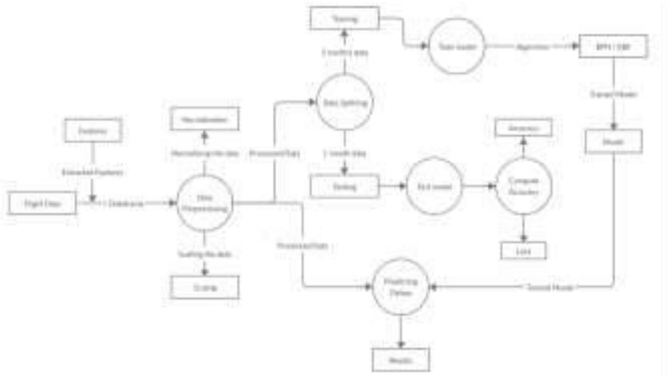

Fig -1: Predictive Model

2.1 Algorithm

2.1.1 Backpropagation Algorithm

Supervised learning is a type of learning which includes Training the machine using data which is having labels that mean some data is already tagged with the correct answer. After that, the machine is feed with a new data algorithm analyses the training data (set of training examples) and produces a correct outcome from labeled data. Backpropagation is a very important part of neural net training. It is the method which includes the weights of a neural net based on the error computed (i.e., iteration).

$$\delta^L = \nabla_a C \odot \sigma'(z^L)$$

$$\delta^l = ((w^{l+1})^T \delta^{l+1}) \odot \sigma'(z^l)$$

$$\frac{\partial C}{\partial b_j^l} = \delta_j^l$$

$$\frac{\partial C}{\partial w_{jk}^l} = a_k^{l-1} \delta_j^l$$

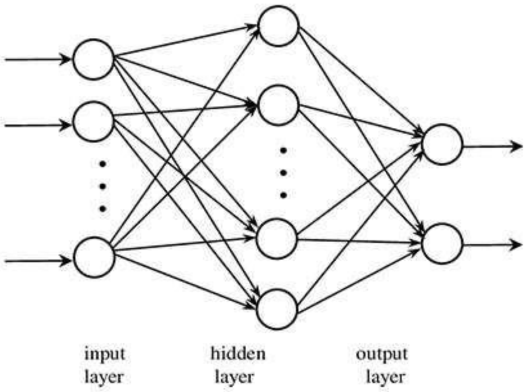

Fig -2: BPN Architecture

Backpropagation works in the following steps,

1. Initially, random weights are assigned to the edges in the neural network.
2. Input nodes are feed with the input values and values of the hidden nodes are computed by Summing the multiplication of input values and weights.
3. This step is repeated until the value of the output node is calculated.
4. Then Error is calculated by getting deference between obtained value and Actual Value.

5. This error is Backpropagated to the hidden layer and using these values, new weights are calculated up to the input layer.
6. These steps are repeated until Value of Error is Minimum.

2.1.2 Radial Basis Function

Unsupervised learning is a type of training of machine that is not having any labels so we use the algorithm to compute on provided information without help. An RBFN classifies the data by measuring the data's similarity samples from the training set. Each RBFN neuron stores neurons centers. When classifying, each neuron calculates the Euclidean distance between the input and its neuron weight. By using this distance data is predicted by computing the cluster which it belongs to.

RBFN can be implemented by Unsupervised as well as semisupervised learning methods.

Radial base works in the following steps:

1. Input layers are feed with the input vector.
2. RBF neurons enter is initialized with the k-means clustering.
3. For every data, its Euclidean Distance is calculated.
4. Finally, the sum of the weights of the output nodes is computed as final output score.

$$f(\mathrm{x}) = \sum_{j=1}^{m} w_j h_j(\mathrm{x})$$

$$h(x) = \exp\left(- \frac{(x-c)^2}{r^2} \right)$$

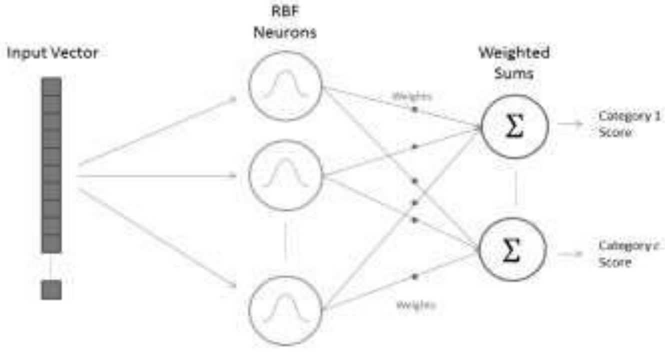

Fig -3: RBF Architecture

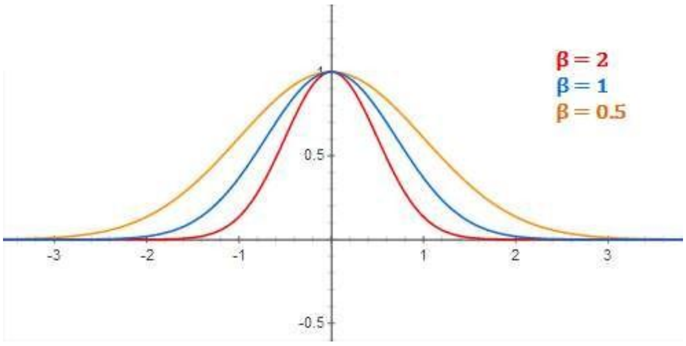

Fig -4: Bell Curve for Beta Values

As we have used the k-means clustering method for selecting the centers for RBF neurons, we can find the value of beta by calculating the sigma as the average distance between all points and their cluster centers.

$$\sigma = \frac{1}{m}\sum_{i=1}^{m} \|x_i - \mu\|$$

Then we computed beta bv the following formula,

$$\beta = \frac{1}{2\sigma^2}$$

2.2 Methodology

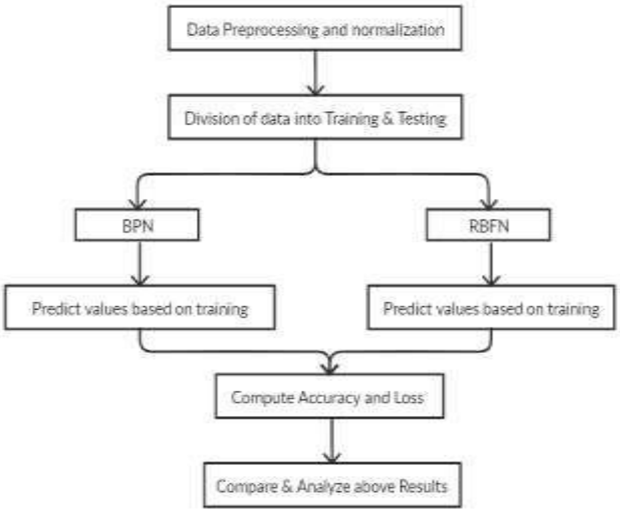
Data Preprocessing and normalization
Division of data into Training & Testing
BPN
RBFN
Predict values based on training
Predict values based on training
Compute Accuracy and Loss
Compare & Analyze above Results

Fig -5: Flowchart of Model

1. Data Preprocessing and Data Normalization : Data analysis and preprocessing is performed on the dataset in order to extract the meanings from it. Normalization is performed by scaling the data to a certain range so that it makes the computation of output efficient.

2. Data Splitting :
Data is divided into training and testing part. Training data is used for training the predictive model and for evaluating models' performance, testing data is used.

3. Model Definition :
Model Definition consists of defining models with certain Hyperparameters. Then the model is trained using the training data.

4. Prediction :
Computing models' accuracy and loss percentage. These values help us in evaluating the performance of the model implemented.

5. Model Evaluation :
Finally, the results are compared and analyzed.

2.3 Analysis

The data will be normalized and scaled so as to use for training and prediction. Data exploration is the initial step in The Analysis of Data, and points of interest. This process isn't meant to reveal to study in greater detail. Data exploration can use data visualizations, charts, and initial reports.

2.3.1 Data Attributes

1. **MONTH, DAY, DAY_OF_WEEK:** dates of the flight
2. **AIRLINE:**
An identification number assigned by US DOT to identify a unique airline
3. **ORIGIN_AIRPORT and DESTINATION_AIRPORT:**
Code attributed to identify the airports
4. **SCHEDULED_DEPARTURE & ARRIVAL :**
Scheduled times of take-off and landing
5. **DEPARTURE_DELAY and ARRIVAL_DELAY:**
Difference (in minutes) between planned and real times.

2.3.2 Data Exploration

Data Exploration is a way in data analysis in which analysts visualization methods for exploring the data and finding meanings in them.

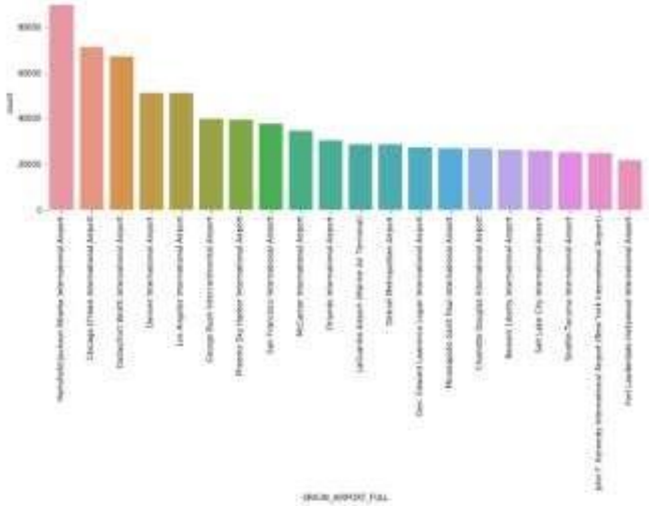

Fig -6: No. of flights with Origin Airport

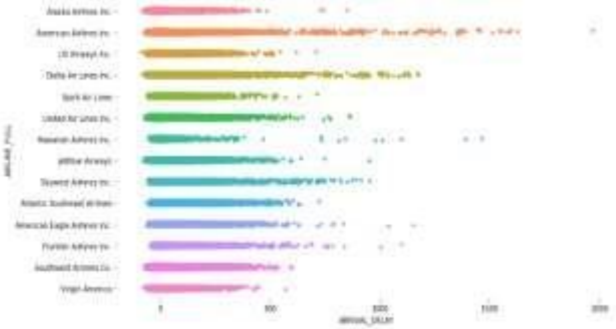

Fig -7: Arrival Delays of Airlines

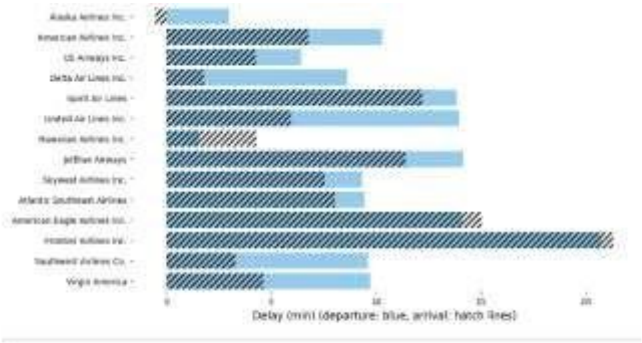

Fig -8: Delay on Departure and Arrival of Flights

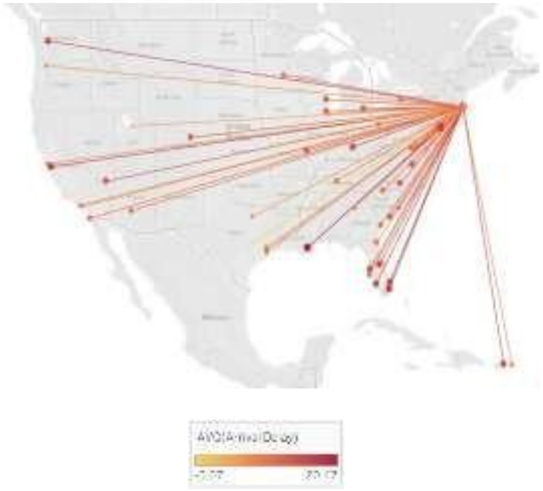

Fig -9: Routes Representing Arrival Delays of Flights

2.3.3 Clustering

The elbow method may be a heuristic method of interpretation and validation of consistency within-cluster analysis designed to assist find the acceptable number of clusters during a dataset. It’s often ambiguous and not very reliable, and hence other approaches for determining the number of clusters like the silhouette method are preferable Explained variance. The "elbow" is indicated by blue dots.

The number of clusters chosen should, therefore, be 6.

This method looks at the percentage of variance explained as a function of the number of clusters: One should choose a number of clusters that give much better modeling of the data. More precisely, but the graph.

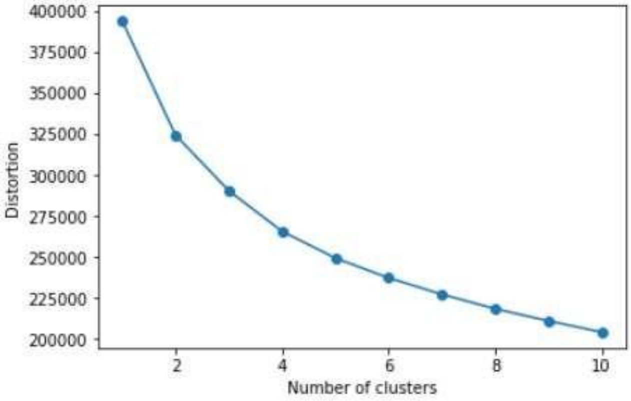

Fig -10: Elbow Method

The K-means clustering algorithm proceeds as follow 1. Set centers of clusters randomly pattern.
2. Calculate Euclidean Distance between points.
3. After calculating all the values, computing centers of clusters by averaging the values
4. Perform steps 2 & 3 until difference does not changes over the iterations.

Table -1: Flight Delay Dataset

Dataset	No. of attributes	No. of instances
2015 Flight Delays and Cancellations	10	1403471 (3 months)

The dataset will be divided into two parts, train and test. 30% of the data will be used for testing and the rest for training. Based on the accuracy of the results the size of the dataset would be further increased.

2.3.4 Visualization

The output computed by the model will be visualized for studying both the results. Visualization will be done using libraries like Matplotlib, Seaborn, etc.

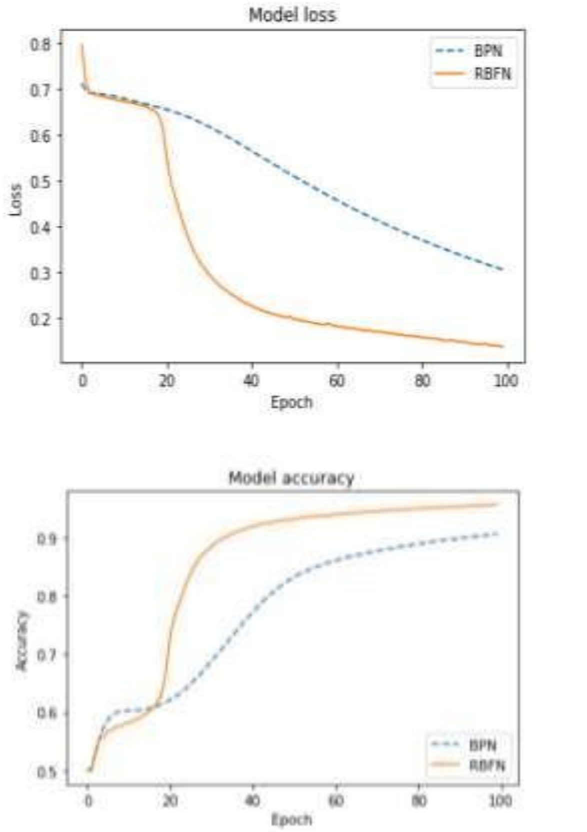

Fig -11: Model Results

2.3.4 Analyzing the Results

Confusion Matrix is The Prediction Results Summary classification problem. The number of correct and incorrect prediction are grouped down by each class. The confusion matrix explains that the classification model is performing well. It gives us insight not only into the errors being made by a being made.

Precision - No. True predictions that were correct.

Recall - /percentage of /true values computed.

ROC curve - An ROC curve is a graph showing the performance of curve plots two parameters:

I) True Positive Rate
II) False Positive Rate

A ROC curve plots TPR vs. FPR at different classification thresholds. Lowering the classification threshold classifies data and True Positives.

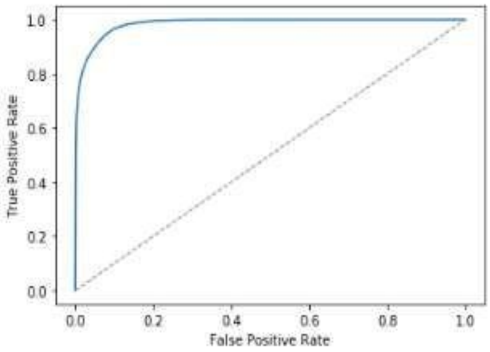

Fig -12: ROC Curve for BPN

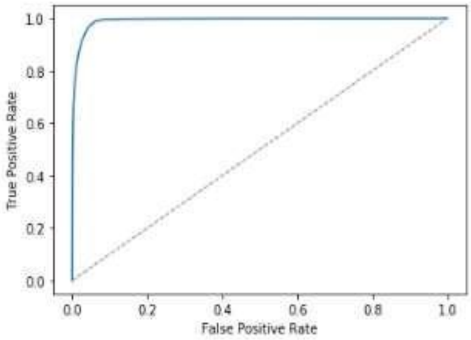

Fig -13: ROC Curve for RBFN

Table -1: Training and Testing Accuracy of Models

Algorithm	Training Accuracy	Testing Accuracy
BPN	90.63%	90.4%
RBF	95.71%	96.05%

Table -1: Comparison of Accuracy of Models

Algorithm	Precision Score	Recall Score
BPN	97.20%	83.3%
RBF	95.93%	96.3%

Based on the results obtained after testing and training, conclusions will be made based on the following factors in order to choose and employ the best Algorithm:

Technique	Accuracy	Epochs	Loss
BPN	90.4%	100	30.4%
RBF	96.05%	100	14.05%

III. CONCLUSION

In this paper, we study the Flight Delay Classification based on neural network techniques. The Objective of this study is to create an effective model i.e. neural models to help us make a proper classification of flight delay. In this system, the implementation of BPN and RBF Neural Network has been proposed. The results conclude that RBFs can be trained much faster than the perceptron. The smallest training error was achieved with RBFN. The Classification taker by BPN is more than RBFN. Future work will include bigger data training and prediction, predicting flight delays as Origin and Destination Airports Delays.

REFERENCES

[1]. C. Charalambous, A.Charitou and F. Kaourou, "Comparative analysis of artificial neural network models: application in bankruptcy prediction", IJCNN'99. International Joint Conference on Neural Networks. Proceedings.

[2]. Nurhakimah Binti Abd Aziz and Wan Fazlida Hanim, "Abdullah Comparison between MLP and RBF network in improving CHEMFET sensor selectivity", 2015 IEEE Symposium on Computer Applications & Industrial Electronics (ISCAIE)

[3]. Bencharef Omar, Bousbaa Zineb and Aida Cortés Jofré, "A Comparative Study of Machine Learning Algorithms for Financial Data Prediction 2018", International Symposium on Advanced Electrical and Communication Technologies.

[4]. S. Vani, T. V. Madhusudhana Rao and Ch. Kannam, "Naidu Comparative Analysis on variants of Neural Networks: An Experimental Study", 2019 5th International

[5]. Angelos P. Markopoulos, Sotirios Georgiopoulos and Dimitrios E. Manolakos, "On the use of Backpropagation and radial basis function neural networks in surface roughness prediction.", Journal of Industrial Engineering International 12, (2016).

[6]. 2015 Flight Delays and Cancellations https://www.kaggle.com/usdot/flight-delays

[7]. Petra Vidnerova. 2019. RBF Layer for Keras Library. https://github.com/PetraVidnerova/rbf_keras.

[8]. Chris McCormick Radial Basis Function Network (RBFN) https://mccormickml.com/2013/08/15/radial-basisfunction-network-rbfn-tutorial