JOURNAL OF CRITICAL REVIEWS

ISSN- 2394-5125 VOL 7, ISSUE 13, 2020

MULTI-NEURAL NETWORK MODEL FOR HANDLING REAL-TIME APPLICATIONS USING EMBEDDED HADOOP CLUSTER

Karthikeyan S[1], Hari Seetha[2]

[1]Assistant Professor, VIT-AP University-Amaravati, Andhra Pradesh-India
[2]Professor, VIT-AP University-Amaravati, Andhra Pradesh-India
[1]karthikeyan.s@vitap.ac.in

Received: April 2020 Revised and Accepted: June 2020

ABSTRACT—Data analytics is a key feature and requirement for the growth of any industry or organization. However, today's small-scale industries with low budgets and big data often face relinquished profit issues, which can only be solved by a combination of big data warehouse systems and efficient data analytics tools. The paper considers the problem posted due to lack of data analytics attached to big database structures or efficient data mining and analytics tools with the absence of effective databases. This paper looks forward to developing a working idea on embedded Hadoop with predictive and statistical data analytic tools like the Artificial, Convolutional and Generative neural networks and machine learning algorithms like Gradient boosting (xgboost), support vector machine (SVM) and regularization regression models. The research takes into consideration the harsh and noise persisting data and use multilevel neural networks to solve this problem. This paper made use of blob segmentation and detection for better identification of hotter regions.

KEYWORDS—Hadoop, neural networks, Machine Learning, ANN, CNN, RNN, GAN, SVM, XGBOOST, Gradient Boosting, Deep Learning, Lasso regression, Ridge regression, Blob Segmentation

I. INTRODUCTION

System design and data analytics have been often differentiated and concentrated in different fields of study. However, we see the existence of various problems exist which require both the understanding and implementation of databases which are large and rapid. Databases which are either SQL or NoSQL can be extracted with ease in order to do fast data preprocessing, data interpretation, hypothesis testing and prediction. However, even with the presence of such superior database structure, there is a lack of cumulative workbench. The term is denoted as the existence of outliers as per statistical terminology is the scenario which is being witnessed in today's technological world, even while you see a great number of problems being solved by NoSQL and SQL databases we can still find a great number of outliers existing outside this pretty NoSQL skewed curve.

Today not only weather forecast has become an issue of the multidisciplinary attribute, unlike back in the 19th century where the computation was a lot easy. Currently, we not only deal with basic features like temperature, humidity, heat maps and global positioned location but also with its historical data which includes its past 5 years' data (due to the effect of global warming), global warming, increase in the depletion rate of ozone layer or other environmental as well as atmospheric variations like increase or decrease in forest cover and many other affecting factors. Weather forecast is just one example of how badly the industry requires both data storing and analyzing streams to merge in and find a solution for weather forecasting, banking transactions, software enterprise data storage and many more aspects in various fields.

The usage of different data preprocessing, cleaning techniques have been used in a way to not only implement the idea of developing a system with enough ability and store good data but also to analyze and adjust to the different scenario of industrial, technical and non-profit problems. Hadoop Cluster is setup as a multi-node, multidimensional stack of storage of data, it also takes away the SQL and NoSQL privileges. However, the utility use of Pig and Hive are still trying to reach a wide range of operations. The structure of Hadoop Distributed File System enabled with jupyter or efficiently calling python script will not only enable us with high ability and efficiency of working with data but also a protocol of entry of data via Jupyter going head by data pre-processing into the Hadoop HDFS cluster will reduce redundancy , null or missing values, and potentially can also work with categorical variables depending on what machine learning algorithm wants to have for the training and testing datasets . Ultimately, the question that looks bright and clear but still persisting in scientific communities for years is how to get the idea from a mind simulation to hardware-software design

ISSN- 2394-5125 VOL 7, ISSUE 13, 2020

and finally to the proclaimed area where the idea was thought of. So, to get it on rails an automated intelligent bot is designed using various IOT and electronics techniques in such a way that it can be deployed in a way to find and retrieve data directly from the deployment site to the Hadoop system.

II. RELATED WORKS

Aasha begum et.al. have proved that multinode cluster reduces time and increases throughput [8]. Big data is playing a vital role in today's world in the field of IT, storage, communications, etc. Large amount of storage is required for distributed file systems, Hadoop Distributed File Systems is doing a good work. Multi Node cluster reduces time and increases [8] but Cluster management is hard: - In the cluster, operations like debugging, distributing software, collection logs etc. are too hard and still single master which requires care and may limit scaling. In a single node cluster Name node, Date node, Job Tracker and Task Tracker all run on the same system ie. on a single instance of Java Virtual Machine (JVM). Multiclusters also requires configuration settings as it works on architecture fashion as in Master-Slave where Name node is master any communication issues between slaves and master would rise an error. Hence we have preferred usage of single node multi cluster.

Aaron Damashek et.al. have demonstrated the detection of specific guns using edge matching and rendering techniques in [7]. The paper shows a where about of a computer vision project. In our presentation model, we have developed a dataset of about 10 thousand images of 7 different types of weapons and a trading time of more than 27 hours to completely train a network of CNN and RNN over to multiple predictions is achieved. This not only provides more accuracy and ROC AUC score but also enhances computer vision with neural networks and deep learning algorithms. Moreover, this extends the classification of only gun samples heading forth to different varieties of weapon samples in order to provide a more robust and powerful classification idea and support to it.

Atmospheric temperature and pressure analysis using support vector machine [3] is a very efficient idea in a way that the boundary of separation will try to formulate the ideas of the worse scenarios and try to form boundary according to the hyper parameter choice. We have depicted the idea of weather forecasting using XGBoost techniques in order to obtain the best analysis using gradient boosted trees. As the SVM accuracy and F1 score is tentatively lower than gradient boosting algorithm moreover we find the absence of outliers in our models however when compared to box plot existence of outliers are seen in the SVM detection algorithm.

Ogunleye et.al. have proved that Enhanced XGBoost can give good results for chronic kidney diseases [10]. It has inspired us to review existing techniques, and propose a similar model USING XGBoost model for heart diseases to find symptoms of diseases and help users by emailing them details.

Janani et al. have compared various techniques and models used for weather forecasting and its techniques which has helped us to make our Delhi smog threat analysis model where we would be predicting smog on few climatic factors. Prediction of weather is a difficult task even for experienced meteorologists because they have to depend on many factors and predict forecast which should be accurate to present as well as the coming scenario. Nowadays, with growing technologies this task is being worked out with the help of AI.

III. IMPLEMENTATION

The ideological implementation deals with multiple different models to enhance the idea of embedding data storage model Hadoop with Data analytical tool of Machine learning and Deep Learning so as to develop a better automated model which would be of great use to present world. Our model consists of data acquisition module developed by us embedding multiple sensors so as to work on real time data and do tasks.

A. Data Bot

The basic idea is of a bot which is automated and will be available inside the household, to serve the purpose of gathering basic data of temperature, humidity, global position and other basic data features. This bot can be controlled either by gestures using gyro sensor or through a mobile application. However, the other organized function was to sustain as an emergency bot to make a call using a one-touch go app which would directly contact the hospital and inform about the location of the emergency. It would always be located with GPS location of the nearby hospital which would be of help to user using it. The bot is also designed such that it can prevent itself from colliding with obstacles inside area.

B. Weather Forecasting

As priory discussed in the introduction how weather forecasting has changed highly and further automation is needed in order to transform better analysis and better prediction in order to save the masses. The algorithm has

ISSN- 2394-5125 VOL 7, ISSUE 13, 2020

a dataset of about 94 features consisting of parameters like global warming effects, historical evidences, locational information, temperature and a few more feature engineering techniques involving calculation of mean, skew measure in order to understand the motion of the point, kurtosis calibrated value so that the tails of the graph can be seen and the normal distribution difference can also be found. By this way the algorithm has been able to identify the potential cyclone near the coastal ranges of the country, give a precise accuracy of about 78-82% by varying algorithms, and parameter tuning.

C. Honey Bee Production Analysis

All over the world the production of honey bee has been decreased in a pretty vague fashion, however there are only few potential reasons to understand why this has happened. The reasons are factual which are lack of resources, investment, clusters of honey bees and a few more boundaries to sustain. The dataset consists of about 198 features collected data over the world through web scraping and kaggle. The hadoop monitoring system is used as a tool to analyze over a cluster of collected data and as per that send a notification email to the faster about what are the necessary species of honey bee, how many different cluster, investment requirements and finally the possible time period to earn profit for the given investment.

D. Home Security - Face detection and recognition

It has been very well seen now in recent years that since figure prints can be captured and used fingerprint scans are vulnerable, however retina scans are pretty expensive for daily household. But if the same security check can be done by a camera that would solve all issues. We have used CNN to recognize the facial features of different members of a family and have trained the model in such a way that it can identify the desired person. Over to it RNN and LSTM network work to remember all those who are frequent user such that the configuration and analysis time is reduced. Moreover, a dual use is made having a weapon scanner as well. So whenever an unknown face with weapon is encountered the direct message contact is given to the head of the family. When we encounter a member with a weapon a flag is updated and held for 12 hours between which for any investigation purposes the information can be used. With a database compatibility we can store the date, time, name, coordinates of the weapon (holding hands), picture of the family member for any assault related cases or enquiry.

E. Terrorist Detection System

The Pathankot attack in India was a clear evidence of how computational unavailability at the security cameras are responsible for delayed security response and casualties in the nation. Variation Auto encoders and Generative Adversarial Neural Networks for artificial image generation has played a vital role in our model to spread the image of early persisting objects over the frame. When a penetrating object is detected, the algorithm checks whether it's speed is greater than a certain threshold at that instance. Once its greater it emails the incumbents about a presence of life threatening weapon with its current location and along with coordinal dimensions determined with the help of computer graphics. It can also presumably trace back upon the projection and find the projectile's initiation point. This can not only enact upon the already present defense system but can save thousands of lives if properly used. A combination of CNN and RNN has been used for tracing and identification of the weapon or object rendering and mathematical computational tools of graphics to check forth the speed of identified object using multi-video processing and MVEC approach.

F. Data Collection Bot

Now, finally we needed a platform to conclude all the ideas developed in a miniature scaled prototype. Using the microcontroller like Arduino Uno, sensors like Wi-Fi module, driver boards, ultrasonic sensors, and stepper motors we tried to build a bot which will compile all the ideas discussed earlier and get into its final shape with its ability to collect data and transmit to device so as to store it in database. Bot is a 4-wheeler prototype vehicle with sensors mentioned above which can be either controlled by gestures using gyro sensor or application-controlled as per your choice. The codes are framed in C++ compiler. Bot has the ultrasonic sensors attached to it to prevent any accident of it with any of the obstacles around. Gesture controlled or app controlled bot runs on the principle of Wireless Fidelity(Wi-Fi) wireless communication. Furthermore, there is a lot of scope to take these ideas to the next level.

G. Delhi Smog Threat Analysis

The national capital Delhi has been under threat since last 5-7 years, many researches and understanding have been put forth to control this issue. Out of the top global warming Delhi constitutes a great amount of damage percentage however the crop burning, sulphate oxides and other mono oxides that constitute a major part in formation of smog. However only if we can predict the amount of incoming smog so as to either create awareness to the incumbents so that measures of spreading anti-pollutants into the environment can be modified. For training the model we gathered chemical analysis data to understood the outliers which would

ISSN- 2394-5125 VOL 7, ISSUE 13, 2020

have either very high or low values from our sample distribution curve. Later, these values were weighted in order to have a predictive analysis. We have used CatBoost in order to enable a gradient boosting algorithm with an accuracy of 84% with hyper parameter tuning. So this statistical and predictive gradient boosting algorithm gives us the prediction and results which are then emailed directly to the accountable incumbent's email, consisting of information regarding how soon the attack of Smog is a threat to Delhi.

IV. RESULTS

To validate our proposed model, we have developed a bot for data collection so as to perform real time application. We have collected our own data for most cases due to which accuracy is less but if we can gather larger database then we can improve accuracy. Further, we propose results of each model.

A. Weapon Analysis

The graph describes different weapon analysis which are each tested about 50-100 times each. Success denotes the percentage in detecting the sample. Precisely for knife the success rate is 74% i.e. out of different sets of knife being here or not the algorithm predicts 74% times correctly whether or not the knife exists in the frame.

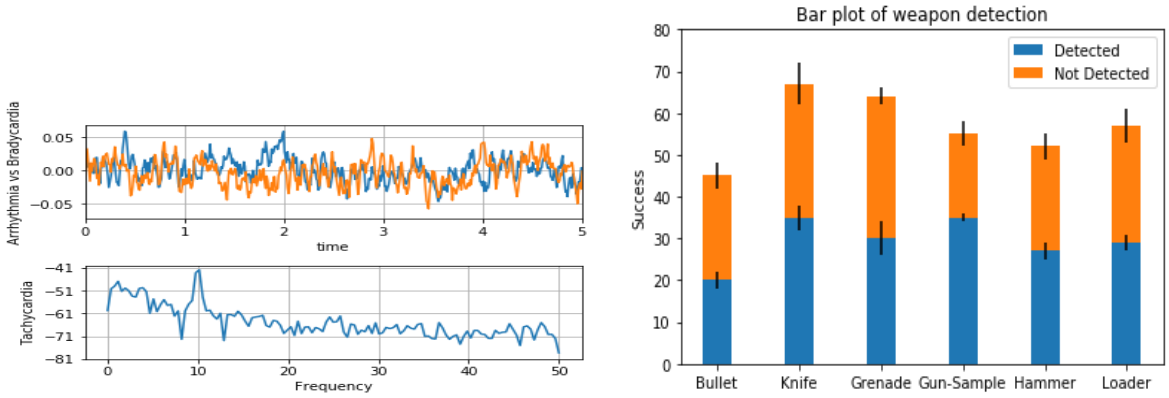

Fig.1. Bar plot for weapon analysis.

B. Health Analysis

The figure is the result of how different heart related disease in a sample house are obtained and plotted with time shows an increment in peaks during the starting few units. However over looking at a frequency which is an inverse of time we observe that in the last couple of units high fall in the slope of the graph hence we can conclude that in a normal household under normal condition if certain heart diseases are present then their majority of growth is seen during the starting time period or precisely the time period when the patient was unaware of the diseases however on longer durations the effects seems to decrease of the trend.

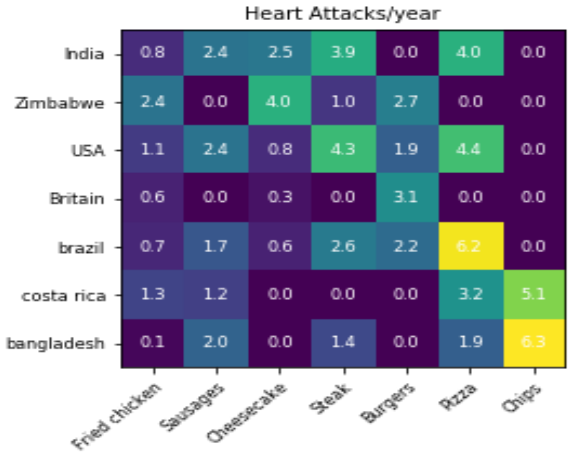

Fig.2. Correlation matrix of health analysis.

This figure 2 indicates the correlation matrix of different countries mapped to different attributes of food (healthy and junk) in a manner of low correlation signifies rarely linked or very less dependent attributes i.e. in

ISSN- 2394-5125 VOL 7, ISSUE 13, 2020

India, heart attack due to intake of chips is not possible or heart attack and chips intake is not at all related however pizza and heart attack is highly related that is 4.0 units or (0.04 of correlation) . Since heart attack is a pretty vague term with many relations as well as hereditary possibilities we have multiplied the correlations before plotting in such a way that the graph can be easily visualized as well as understanding can be made easy.

C. Delhi Smog Analysis

The figure shows the error analysis in the prediction of Delhi smog during the winters as a box plot and Notch plot. It demonstrates the presence of high outliers in central Delhi, the average air pollutant quality of central Delhi is too above from that of south and north Delhi because of the presence of industries and high vehicle densities as per general understanding.

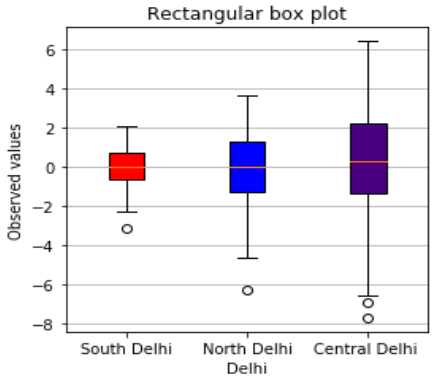

Fig.3. Box plot for smog analysis.

Thus in the model we have ensured we remove these outliers so that the prediction can be generalized as well as correct by using XGBoost classifier to about 200 trees that will weighted to decrease the vote in such a way that the outlier positive prediction won't take place. This graph or radar analysis shows the entry of smog trends where red dots show high possibility of pollutants and light blue the least. Green dots show pollutants which are very less harmful and can be tolerated, yellow dots are bio pollutants.

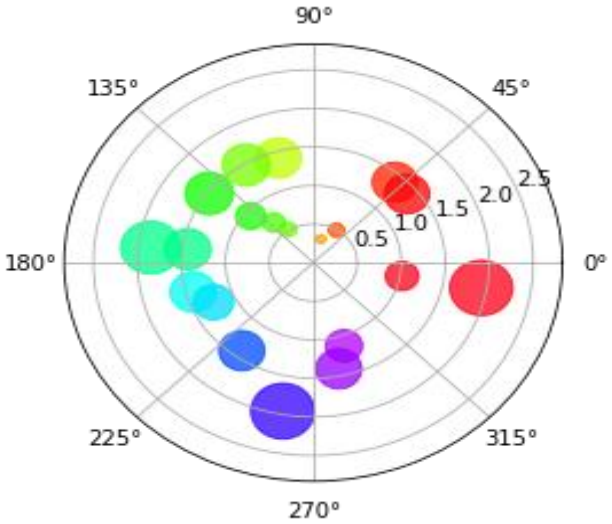

Fig.4. Statistical analysis.

D. Statistical Analysis

The measured statistical report of Honey Bee production accuracy, Delhi smog analysis, weather forecasting and cyclone prediction are given below in the respective order. The high F1 score as well as accuracy suggest the dependability of the model .

V. CONCLUSION

In this paper, proposed the requirement and benefits of embedding optimized databases and data analytics tool. Our model shows how effectively the System Design using tools like Hadoop which can be used to solve data issues, other tools make analyses such as weather analysis, Heart attack and heart health analysis, honey bee production and bot for data collection. The overall accuracies for all machine learning classification algorithms

of our model ranges from 83 to 87 percent with a standard deviation of 7.96. However, the image recognition and deep learning models namely GAN's, CNN and RNN have pretty low accuracies due to lack of multiple cases dataset of about 73 to 75% but that can be enhanced using a larger database with more cases. The methods and results prove the requirement of system design when compared to existing individual approaches. For all classification we have used Gradient Boosted Tree so that error learning can also be used in an efficient way. Embedded Hadoop data analysis in multilevel neural system for various data issues

VI. REFERENCES

[1] L. Prokhorenkova (Ostroumova)G. GusevA. VorobevA. DorogushA. Gulin, CatBoost: unbiased boosting with categorical features, NeurlPS,2018.

[2] Janani.B , Priyanka Sebastian, ANALYSIS ON THE WEATHER FORECASTING AND TECHNIQUES, IJARCET, Volume 3, January 2014.

[3] Y.Radhika And M. Shashi, "Atmospheric Temperature Prediction Using Support Vector Machines," International Journal Of Computer Theory And Engineering,vol.1,no.1,Apr 2009.

[4] Dr.S.Santhosh Baboo And I.Kadar Shereef, "An Efficient Weather Forecasting System using ANN" International Journal Of Environment science and development,vol.1,no.4,oct 2010.

[5] Bhavna Khajone, Prof. V. K. Shandilya, Concealed Weapon Detection Using Image Processing, International Journal of Scientific & Engineering Research, Volume 3, Issue 6, June-2012.

[6] Justin Lai, Sydney Maples, Developing a Real-Time Gun Detection Classifier, Course: CS231n, Stanford University,2017.

[7] Aaron Damashek and John Doherty Detecting guns using parametric edge matching Project for Computer Vision Course: CS231A, Stanford University, 2015.

[8] A. Aashabegum and K. Chitra, "Formation of Single and Multi-Node Clusters in Hadoop Distributed File System," *2017 World Congress on Computing and Communication Technologies (WCCCT)*, Tiruchirappalli, 2017, pp. 162-164

[9] C. Verma and R. Pandey, "Big Data representation for grade analysis through Hadoop framework," *2016 6th International Conference - Cloud System and Big Data Engineering (Confluence)*, Noida, 2016, pp. 312-315.

[10] A. Ogunleye and Q. Wang, "Enhanced XGBoost-Based Automatic Diagnosis System for Chronic Kidney Disease," *2018 IEEE 14th International Conference on Control and Automation (ICCA)*, Anchorage, AK, 2018, pp. 805-810.

[11] Kamini, Silky Sachar, Sonia Suneja(2017), "Review Paper on Mean Stack for Web Development", International Journal for Scientific Research & Development;Vol. 5, Issue 01; ISSN (online): 2321-0613,_pg;497-498.