A Novel Approach to Predict Chronic Kidney Disease using Machine Learning Algorithms

Bhavya Gudeti
School of Computer Science and Engineering
Vellore Institute of Technology,
Chennai, 600127, Tamilnadu, India.
bhavyagudeti@gmail.com

Shashvi Mishra
School of Computer Science and Engineering
Vellore Institute of Technology,
Chennai, 600127,
Tamilnadu, India.
shashvimishra@gmail.com

Shaveta Malik
Terna College of Engineering,
University of Mumbai ,
Maharashtra, India
shavetamalik687@gmail.com

Terrance Frederick Fernandez
*[0000-0002-7317-3362]**
Rajiv Gandhi College of Engineering and Technology, Puducherry, India
frederick@pec.edu

Amit Kumar Tyagi *[0000-0003-2657-8700]**
School of Computer Science and Engineering,
Vellore Institute of Technology,
Chennai, 600127, Tamilnadu, India.
amitkrtyagi025@gmail.com

Shabnam Kumari
Anumit Academy of Research and Research Network, India
shabnam.kt25@gmail.com

***Abstract*—A staggering 63,538 cases have been registered, according to India's health statistics upon Chronic Kidney Disease (CKD). The average age of nephropathy for humans lies between 48-70 years. CKD is more prevalent among males than females. Bitterly, our Nation rank among top 17 countries in CKD since 2015 which is characterized by a gradual loss of excretory organ performance over time. Earlier detection of the illness followed by treatment could keep this dreaded disease at the shore. Machine Learning, is making sensible applications in areas such as analyzing medical science outcomes, sleuthing fraud etc. For the prediction of chronic diseases various machine learning algorithms are implemented.**

Our main aim is to differentiate the performance of various machine learning algorithms primarily based on its accuracy. In this work we idolized Rcode to compare their performances. The pivotal purpose of this study is to analyze the Chronic Kidney Disease dataset and conduct CKD and Non CKD classification cases.

Keywords— MachineLearning, Chronic Kidney Disease, Classification, Accuracy, LogisticRegression, Support Vector Machine

I. INTRODUCTION

Way back in 1950s, the communication among the human were predominantly oral. However as technology progressed since, mankind were obsessed with the technology. The million dollar query remains, "*Why Humans are more obsessed with technology?".* The response is straight-forward. Demand rise in manufacturing accelerates surplus data in trade growth, product, business perspective and sales. These days industries like automation, aerospace, health care, etc., are operating in communication of Machines or interconnection of Internet of Things (IoTs). These IoTs devices (in interconnection) are manufacturing heaps of knowledge that is required to be analyzed with efficiency through efficient and fashionable tools/approaches. Current available ancient tools don't seem to be enough to analyze huge volumes of data.

Clustering may be thought as an assortment of objects into clusters that are similar in nature. The cluster/group contains objects that mimic each other, while the objects in the other ones are dissimilar. The application is widely applicable in applications like Marketing, World Wide Web (WWW), Earthquake Studies, Aerospace, Biology, Insurance, etc. On another hand, if the information renders with categorizes/ class labels, and then classification technique is employed to categorize the given information into number of classes/ categories based on their similarities. The various applications of classification are speech and handwriting recognition, Identification of biometric, classification of documents, etc. Association Rule Mining (ARM) is: if-then statements facilitate to indicate the relationships between data items among transactional databases. Further, Regression (or linear regression) is employed to seek out the relationship between two continuous variables. One variable is termed as predictor or independent variable and other is dependent or response variable. Outlier detection is outlined as "The method of distinguishing the extreme points within the data. It is a branch of data mining." These all algorithms (discussed above) are a part of Data mining/ Machine Learning/ Computer Vision.

In the human body, the kidney is instrumental in absorbing and discharging all the toxic and unessential materials, typically wastes, from the body through egesting and excretion process. In India, there are approximately one million cases of Chronic Kidney Disease (CKD) every year. It is dangerous to kidney and it produces gradual loss in kidney functionality. Nevertheless, it is unpredictable because its symptoms grow gradually and are not unique to the disorder, it is important to detect CKD at its early stage. Kidneys filter wastes and excess fluids from the blood that are then excreted in excrement. In the early stages of CKD, we will have a few signs or symptoms.

In healthcare organization, Classification is one in all the foremost usually used ways of machine learning. The

classification model shows the class of result for each data point.

The classifying methods are Decision Tree, Support Vector Machine, K-Nearest Neighbor, Naïve Bayes classifier, and Neural Networks. KNN is used to visualize at the relationship between different CKD risk factors, in order to predict the disease at an early stage. Machine Learning is a growing phase dealing with the study of a huge variable data and it is grown from the study of pattern (speech and handwriting) recognition and computational learning theory in Artificial Intelligence having numerous methods, algorithms, and techniques to analyze and predict the data. Machine Learning techniques have proved huge success in detection and recognition of many essential diseases in medical science's point of view. Machine learning would thus be useful for predicting whether the patient has CKD or not in this question. By using old CKD patient data to train predictive model, Machine Learning does so.

A. Analysis of Chronic Kidney Disease (CKD)

It makes its way as a ground-breaking and actual channel that liberates the body from squander and parlous substances and return supplements, amino acids, insulin, hormones and different basic substances to the circulatory framework. Incidentally things will flip out gravely, however. "Chronic Kidney Disease (CKD) is used at some stage in the world to suggest to any variety of nephritis that returns. "Infection" incorporates any deviation from the urinary organ structure or limit customary, paying very little heed to whether or not it is most likely going to create a man feel unwell or manufacture complexities. It is a typical issue that may influence anybody at any age. It's assessed that just about 3 million people within the United Kingdom are at risk of CKD. A combination of totally different conditions that usually place as train on the kidneys works on CKD.

Hence, the manuscript is organized as follows: Section II mentions about related work upon this research topic. Section III, discusses the Proposed System to classify Chronic Kidney Disease (CKD). Section IV describes the information regarding the dataset used and transient introduction regarding the attributes. Section V deals with the machine learning algorithms, code and its results for variable measures and therefore the corresponding output obtained in each classification algorithm. Further, section VI discusses about an open discussion about current view, results about chronic disease. Section VII finally discusses the conclusion of the research work alongside with the attribute improvement.

II. `RELATED WORK

There are diverse researchers who have worked with the assistance of several different classification algorithms on CKD prediction. All those had their model performance expected. Gunarathne, W.H.S.D. [1] compared the effects of divergent models. Finally, they concluded that the Multiclass Decision forest algorithm provides plentiful precision for the 14-attribute (reduced) data set. S.Dilli Arasu and Dr. R. Thirumalaiselvi [2] worked on missing values in a Chronic Kidney Disease dataset. They deduced that the missing values in the dataset can not only reduce the model's accuracy but also the effects of the prediction. By patterning a numerical method on stages of Chronic Kidney Disease, they found a solution to this issue and by doing so; they stood up with unknown values. They substituted the missing values with those recalculated ones.

In discovering Chronic Kidney Disease using machine learning algorithms, Asif Salekin and John stankovic[3] used novel approach. They got findings on a dataset consisting of 400 records and 25 attributes resulting in a patient prone to CKD or not. In order to achieve results, they used KNN, random forest and Neural Network algorithms. They used wrapper methodology for feature reduction which finds CKD with high accuracy.

12 different classification algorithms on various datasets were tested by Sahil Sharma, Vinod Sharma, and Atul Sharma [4], each with 400 records and 24 attributes. They compared their expected outcomes with actual results in order to determine predictive accuracy. They used metrics such as precision, sensitivity, accuracy and specificity for measuring the performance of the classifiers. Note that Chronic Kidney Disease (CKD) is not uncommon.

However, a lot of correct information regarding risk for progression to nephropathy is direly needed for clinical selections concerning testing, treatment and referral. Hence, this section highlighted upon the state of art in the field of CKD. Interestingly, the further section would discuss our work in detail.

III. PROPOSED SYSTEM

The proposed system deals with the detection of Chronic Kidney disease. The healthcare systems generate colossal data. Thus, it is obligatory to use this data productively to analyze, predict, and to treat an explicit disease. A classification model offers some solution from determined values. In classification type, we have a tendency to expect fewer or lots of input to predict values of their outcomes. In a supervised machine learning algorithms, the classification algorithm uses the training dataset. Classification predicts the categorical class labels in the data.

The research work tries to present a machine learning framework for information discovery on the Chronic Kidney Disease dataset. To classify the disease at puerility, three machine learning algorithms are used, namely Logistic Regression, Support Vector Machine, and K-Nearest Neighbors. The molarity of every algorithm is inspected. Our proposed model combines the Support Vector Machine, Logistic Regression and K-Nearest Neighbours (KNN) as mentioned in Fig.1.

This section snapshot our system that has been proposed in this paper. Now, next section will discuss the datasets that are being employed and introduce the table that shows the attributes and description on the same.

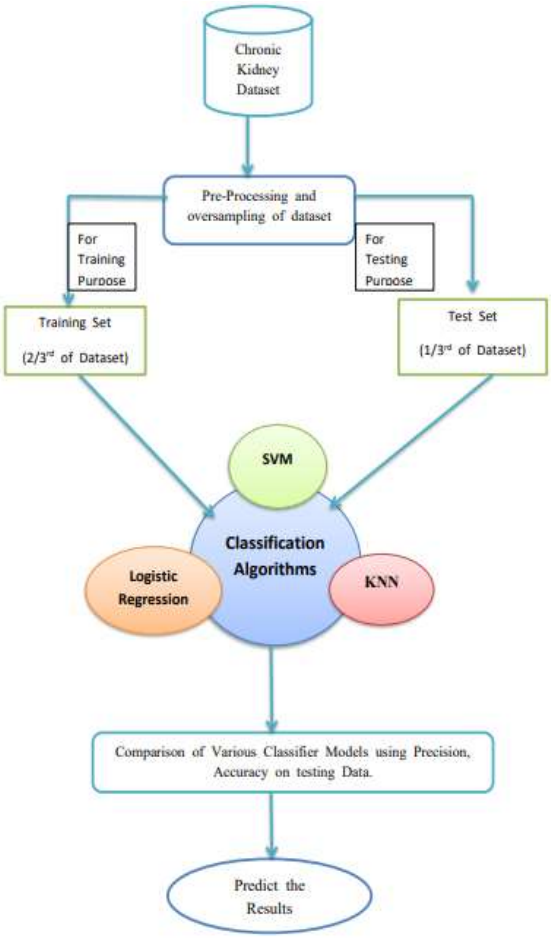

Fig.1. Proposed Model using Various Classification Algorithms

IV. `DATASET USED

The proposed framework uses the UCI Machine Learning Repository dataset called Chronic Kidney Disease (CKD) that has 25 attributes, out of which, 11 are numerical and 14 are nominal. Entire 400 instances of the dataset are used for training to predict machine learning algorithms. In 400 instances, 250 are labeled as Chronic Kidney Disease (CKD) and 150 are labeled as Non Chronic Kidney disease. The attributes present in the data set are bacteria, sodium, age, Hemoglobin, Diabetes Mellitus, Classification, Appetite, Coronary Artery Disease, Blood Pressure, Pus cell, Anemia, Pedal Edema, Sugar, White Blood Cell Count, Hypertension, Red Blood Cell Count, Potassium, Specific Gravity, Pus cell chumps, Packed Cell Volume, Albumin, Serum Creatinine, Red Blood Cells, Blood Urea, and Blood Glucose Random.

The dataset that is taken is divided into two groups, one for testing the samples and another for training the samples. The ratio for testing and training data is 30% and 70% respectively. The data set used has been listed in table 1. The readers can refer following URL [16] for collecting data. Now, next section will discuss regarding the machine learning algorithms used to classify CKD.

TABLE 1 :DATA SET USED

S.No	Attribute	Description about the attribute
1.	Bacteria(nominal)	ba – (present / not present)
2.	Sodium(numerical)	sod in mEq/L
3.	Age (numerical)	Person'sAgeinYears
4.	Haemoglobin (numerical)	Hemo in grams
5.	Diabetes Mellitus (nominal)	dm – (yes / no)
6.	Class (nominal)	class – (ckd / notckd)
7.	Appetite (nominal)	appet – (good / poor)
8.	Coronary Artery Disease (nominal)	CAD – (yes / no)
9.	Blood Pressure (numerical)	BP in mm/Hg
10.	Pus cell (nominal)	PC – (normal / abnormal)
11.	Anemia (nominal)	ane – (yes / no)
12.	Pedal Edema (nominal)	pe – (yes / no)
13.	Sugar (nominal)	su – (0/1/2/3/4/5)
14.	White Blood CellCount (numerical)	Wc in cells/cumm
15.	Hypertension (nominal)	htn – (yes/no)
16.	Red Blood Cell Count (numerical)	Rc in cells/cumm
17.	Potassium (numerical)	Pot in mEq/L
18.	Specific Gravity (nominal)	Sg - (1.005/1.010/1.015/1.020/1.025)
19.	Pus Cell clumps (nominal)	pcc – (present / notpresent)
20.	Packed Cell Volume (numerical)	P cv
21.	Albumin (nominal)	al – (0/1/2/3/4/5)
22.	Serum Creatinine(numerical)	Sc in mgs/dl
23.	Red Blood Cells (nominal)	RBC – (normal/ abnormal)
24.	Blood Urea (numerical)	Bu in mgs/dl
25.	Blood Glucose Random (numerical)	BGR in mgs/dl

V. SIMULATION RESULTS

This section describes about the simulation results that are being used in the paper here.

A. Logistic Regression

Logistic Regression may be a calculation for order. As per heaps of autonomous factors, the logic is 1/0, Yes/No, True/False. It can be employed to access a paired answer. We have a tendency to utilize the likelihood log as an impoverished variable. Logistic Regression is used for the classification problems in Machine Learning Algorithms. It is a prophetic analysis algorithm and it is based mostly on the concept of probability. It means that, given a certain factor, logistic regression is used to predict an outcome that has two values. The source code is exemplified in Table I and the output in Fig.2. From them, we can deduce that the accuracy of Logistic Regression is 0.7725

TABLE II : RCODE FOR LOGISTIC REGRESSION

```
ckd<- read.csv("C:/Users/bhavya/Desktop/ckd.csv")
ckd
ckd$Type<- NULL
head(ckd)
dim(ckd)
summary(ckd)
names(ckd)
contrasts(ckd$classification)
#Logistic Regression
glm.fit=glm(classification~age+bp+pcv+bu,
data=ckd,family=binomial)
summary(glm.fit)
#predict provides a vector of fitted probabilities.
glm.probab=predict(glm.fit,type="response")
glm.probab[1:20]
glm.predc=rep("ckd",400)
glm.predc[glm.probab>.5]="notckd"
table(glm.predc,ckd$classification)
mean(glm.pred==ckd$classification)
```

```
> glm.pred=rep("ckd",400)
> glm.pred[glm.probs>.5]="notckd"
> table(glm.pred,ckd$classification)

glm.pred ckd ckd\t notckd
  ckd    200    1     41
  notckd  48    1    109
> mean(glm.pred==ckd$classification)
[1] 0.7725
```

Fig.2. Output for Logistic Regression

B. *Support Vector Machines (SVM)*

For each relapse and grouping undertakings, Support Vector Machine, curtailed as SVM, will be used. Multitude of researchers favors it deeply as it provides unbelievable accuracy with less power of activity. In ML, SVM support vector systems are supervised models compatible with learning. Support Vector Machine (SVM) offers a dual platform for regression and classification. This can be used to solve both linear problems and non-linear ones. This algorithm uses a hyper plane to categorize the data points. Within this SVM algorithm, each data point will be plotted as a point in n-dimensional space, with a value of each attribute being the value of a given coordinate. Classification can be accomplished by searching for the right hyper-plane which basically distinguishes between the two CKD and not CKD groups. Table III presents the code behind SVM and from the results in Fig.3, we can witness that the accuracy of SVM = 0.9925187

TABLE III : RCODE FOR SVM

```
#Generate a random number that is 70% of the total number of
rows in dataset.
ckd1 <- sample(1:nrow(ckd),0.7*nrow(ckd))
ckd.train<- ckd[ckd1,]
ckd.test<- ckd[-ckd1,]
set.seed(1)
ckd<-ckd[1:200,]
x=cbind.data.frame(ckd.train[,9:13])
y=ckd.train$classification
dataset=data.frame(x=x, y=as.factor(y))
library(e1071)## Support Vector Machine
svmfit=svm(y~., data=dataset, kernel="radial",gamma=1,
cost=1)
summary(svmfit)
svm.probs=predict(svmfit,type="response")
svm.probs[1:400]
svm.pred=rep("ckd",400)
svm.pred[svm.probs="notckd"]="notckd"
mean(svm.pred==ckd$classification)
```

```
> svm.pred=rep("ckd",400)
> svm.pred[svm.probs="notckd"]="notckd"
> mean(svm.pred==ckd$classification)
[1] 0.9925187
```

Fig.3. Output for SVM

C. *K-Nearest Neighbors Classification:*

The sole performance of the K nearest neighbor classifier algorithm is to predict the target variable by capturing the nearest neighbor class. The nearest class will be known as the target variable using the distance measures like Euclidean distance.

Algorithm:

1. Initialize the parameter K.
2. Calculate the distance between the test sample and all the training samples
3. Sort the distance in the ascending order.
4. Take K-nearest neighbors.
5. Gather the class of the nearest neighbor.
6. Here as we can see the accuracy in KNN = 0.7875

From the algorithm mentioned above, it is evident that the results are better in Support Vector Machine. We provide result with an accuracy of 0.9925187. Now, the subsequent section will provide a conclusion regarding this work in brief adding some future enhancement possibilities with this work.

TABLE IV : SIMULATION RESULTS

Name of Classifier	Accuracy
Logistic Regression	0.7725
Support Vector Machine (SVM)	0.9925187
K-Nearest Neighbour	0.7875

`VI. AN OPEN DISCUSSION

Each classifier's results were evaluated using different evaluation parameters, and cross-checked against over-fitting with 10-fold cross-validation. The technique of nested cross-validation has also helped to fine-tune the model parameters. The tests will be carried out using the Python 3.3 programming language through the Jupyter Notebook web application. Several Sciket-learning libraries were used, which is a free machine learning system platform for Python. Accuracy using F1-measurement, sensitivity, specificity and Area under Curve (AUC) are the assessment measures considered in this analysis. Each model produces different outputs; depending on its parameter values.Thus with the GB model we achieve the best efficiency in detection. This result is better than the results obtained by using a multilayer perceptron algorithm (MLP) single point split, seven attributes, and a 98.4 percent F1 measurement. By contrast, a 98.0 per cent F1-measure was obtained with better efficiency relative to study using RF and five apps.

Some limitations on the dataset used are, however, important to this analysis. Second, the sample size (400 instances) is expected to be low and may affect the reliability of the studies. Second, problem identification is another dataset which has the same features to assess the performance of the data sets. Also, the readers are suggested to read [17, 18, 19, 20, 21 and 22] to know more about artificial intelligence, machine learning and deep learning techniques, i.e., their scope in near future.

VII. Conclusion And Future Work

Aimed to diagnose Chronic Kidney Disease (CKD) at an earlier stage, this manuscript introduced a variety of machine learning algorithms. The models obtained from CKD patients are trained and authenticated with the mentioned input parameters. Support Vector Machine, Logistic Regression and *knn* are analyzed to conduct the study of CKD. The performances of those algorithms were determined primarily on the basis of precision. Our results exemplified that the Support Vector Machine algorithm predicts Chronic Kidney Disease better than Logistic Regression and K-Nearest Neighbors within the narrow limits of this medical scenario.

The benefit of this approach is that the prediction process takes far less time and helps doctors to initiate treatment at the earliest for patients with CKD and further to classify larger population of patients within shorter span. Because the dataset used in this paper is tiny with 400 examples, we prefer to work with larger datasets in the future or compare the results of this dataset with a different dataset with the same. In addition, to help minimise the incidence of CKD, we try to predict if a person with this syndrome chances chronic risk factors such as hypertension, family history of kidney failure and diabetes using the appropriate dataset.

Author's Contributions

Shashvi and Bhavya have drafted this manuscript. Amit Kumar Tyagi and Terrance Frederick Fernandez have analyzed and approved this manuscript for publication.

Conflict of Interest

The authors do not have any conflict concerning publication of this manuscript.

References

[1] L. Rubini, "Early stage of chronic kidney disease UCI machine learning repository,"2015. [Online]. Available:http://archive.ics.uci.edu/ml/datasets/Chronic Kidney Disease.

[2] Asif Salekin, John Stankovic, "Detection of Chronic Kidney Disease and Selectiing Important Predictive Attributes," Proc. IEEE International Conference on Healthcare Informatics (ICHI), IEEE, Oct. 2016, doi:10.1109/ICHI.2016.36.

[3] Q. Zhang and D. Rothenbacher, "Prevalence of chronic kidney disease in population-based studies: systematic review," BMC Public Health, vol. 8, (1), pp. 117, 2008.

[4] K. A. Padmanaban and G. Parthiban, "Applying Machine Learning Techniques for Predicting the Risk of Chronic Kidney Disease," Indian Journal of Science and Technology, vol. 9, (29), 2016.

[5] J. Xiao et al, "Comparison and development of machine learning tools in the prediction of chronic kidney disease progression," Journal of Translational Medicine, vol. 17, (1), pp. 119, 2019.

[6] Sahil Sharma, Vinod Sharma, Atul Sharma, "Performance Based Evaluation of Various Machine Learning Classification Techniques for Chronic Kidney Disease Diagnosis," July18, 2016.

[7] GunarathneW.H.S.D, Perera K.D.M, Kahandawaarachchi K.A.D.C.P, "Performance Evaluation on Machine Learning Classification Techniques for Disease Classification and Forecasting through Data Analytics for Chronic Kidney Disease (CKD)", 2017 IEEE 17thInternational Conference on Bioinformatics and Bioengineering.

[8] S.Ramya, Dr.N.Radha, "Diagnosis of Chronic Kidney Disease Using Machine Learning Algorithms," Proc. International Journal of Innovative Research in Computer and Communication Engineering, Vol. 4, Issue 1, January 2016.

[9] S. A. Shinde and P. R. Rajeswari, "Intelligent health risk prediction systems using machine learning: a review," IJET, vol. 7, no. 3, pp. 1019– 1023, 2018.

[10] A.J. Aljaaf et al, "Early prediction of chronic renal disorder mistreatment machine learning supported by prognosticative analytics," in 2018 IEEE Congress on organic process Computation (CEC), 2018.

[11] J.Xiao et al, "Comparison and development of machine learning tools in the prediction of chronic renal disorder progression," Journal of Translational drugs, vol. 17, (1), pp. 119, 2019.

[12] P. Yang et al, "A review of ensemble strategies in bioinformatics, "Current Bioinformatics, vol. 5, (4), pp. 296-308, 2010.

[13] L.Deng et al, "Prediction of protein-protein interaction sites mistreatment associate ensemble methodology," BMC Bioinformatics, vol. 10, (1), pp. 426, 2009.

[14] M. Moslem and M. Pasha, "Survey of machine learning algorithms fordisease diagnostic," Journal of Intelligent Learning Systems andApplications, vol. 9, (01), pp. 1, 2017.

[15] S.Karamizadeh et al, "Advantage and disadvantage of support vector machine practicality," in 2014 International Conference on laptop, Communications, and management Technology (I4CT), 2014.

[16] http://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease

[17] Akshara Pramod, Harsh Sankar Naicker, Amit Kumar Tyagi, "Machine Learning and Deep Learning: Open Issues and Future Research Directions for Next Ten Years", Book: Computational Analysis and Understanding of Deep Learning for Medical Care: Principles, Methods, and Applications, 2020, Wiley Scrivener, 2020.

[18] Tyagi, Amit Kumar and G, Rekha, Machine Learning with Big Data (March 20, 2019). Proceedings of International Conference on Sustainable Computing in Science, Technology and Management (SUSCOM), Amity University Rajasthan, Jaipur - India, February 26-28, 2019.

[19] Amit Kumar Tyagi, Poonam Chahal, "Artificial Intelligence and Machine Learning Algorithms", Book: Challenges and Applications for Implementing Machine Learning in Computer Vision, IGI Global, 2020.

[20] Amit Kumar Tyagi, G. Rekha, “Challenges of Applying Deep Learning in Real-World Applications”, Book: Challenges and Applications for Implementing Machine Learning in Computer Vision, IGI Global 2020, p. 92-118.

[21] Terrance Frederick Fernandez and M. Pradeep, "Multi-level Predictive with Training Framework (MP with TF) for ranking machine learning algorithms", IEEE proceeding of 4th International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud (I-SMAC 2020), pp.697-703, ISBN: 978-1-7281-5464-0/20, 7th to 9th October 2020, SCAD Palladum, Tamil Nadu.

[22] Aravindan C, Terrance Frederick Fernandez, Hema Malini V and Catherine Madhu Vidha J, “An Extensive Research on Cyber Threats using Learning Algorithm”, IEEE proceeding of International Conference on Emerging Trends in Information Technology and Engineering, ISBN: 978-1-7281-4141-1, 25th February 2020, Vellore, India.