Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 16780 - 16790
Received 20 May 2021; Accepted 3 June 2021.

Performance Analysis Of Hyperpipes In The Classification Of Large Datasets

[1*]Subba Rao Peram, [2]B Tarakeswara Rao, [3]Koduganti Venkata Rao, [4]B Premamayudu

[1,4] Department of Information Technology, Vignan's Foundation for Science Technology & Research (Deemed to be University), Vadlamudi, Guntur, AP-India.
[2]Department of Computer Science and Engineering, Kallam Haranadhareddy Institute of Technology, Guntur, AP-India
[3]Department of Computer Science and Engineering, Vignan's Institute of Information Technology, Gajuwaka, Visakhaptnam (Dt),Andhra Pradesh, India

ABSTRACT

Data mining is a study of identifying and obtaining the necessary knowledge from the huge amount of data. Classification is a technique that describes the model by finding the unknown class label of the data object. This paper represents Hyperpipes as a simple technique used for faster classification. The aim of this paper is to analyze the performance of Hyperpipes on different large data sets. For this study, five datasets have been taken from publicly available UCI machine learning repository web site. Results show that this model can be efficiently utilized in the classification of huge datasets which contains a large number of attributes and instances. This model achieves a better accuracy of 98.95% by overcoming the maximum accuracy of 84.15% and minimum accuracy of 39.23% from the data sets.

Index terms
Data mining, Classification, Hyperpipes, UCI, Performance.

I Introduction

Data mining is an art of developing interesting patterns and mining some useful knowledge from the huge amount of data. This Data mining is also known as Knowledge Discovery Process. The hidden knowledge in the large amount of data is extracted from this Knowledge Discovery Process. The value of data has improved with the development of some data discovery techniques. There are different types of methods which are already in use to extract patterns from the stored data which are valuable and previously unknown. This type of information can be Predictive or Descriptive. The basic task of this process is to obtain information from low level databases. Data mining methods are applied for extracting the knowledge from raw data in KDD process[26]. This KDD process is defined in multiple steps that help in converting of data into useful information. The steps in Knowledge Discovery Process are described as follows.

1. **Selection:** In this step, the dataset is selected and the domain is recognized from the data which is to be extracted.
2. **Preprocessing:** During this stage, the selected data may contain some noise and missing values. This type of data may create outliers or irrelevant data and should be removed by cleaning.
3. **Transformation:** In this phase, the nominal data are converted into numerical and are standardized by defining new attributes. Here at this time, the data is managed by various data mining techniques like classification, regression and clustering.
4. **Data Mining:** Data mining is an art of developing interesting patterns and mining some useful knowledge from the huge amount of data.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 16780 - 16790
Received 20 May 2021; Accepted 3 June 2021.

5. **Interpretation and Evaluation:** This is the final phase that consists of results obtained from previous steps. These results should be desperately investigated and competed with previously obtained knowledge.

The following diagram shows the process of Knowledge discovery process.

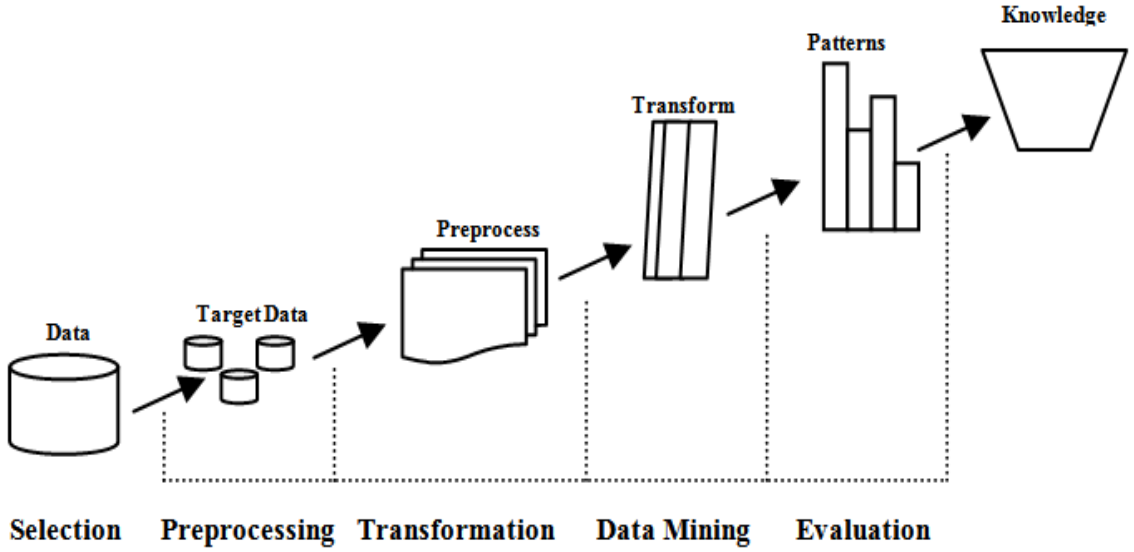

Fig.1. Representation of Knowledge Discovery Process

There is no correct algorithm in data mining for applying on a dataset. Many composite models may suitable, but they are also being problematic for understanding the data. For easy translation, many of the successful applications uses simple models and every model turns itself to an appropriate type of problem. Data mining generates results by applying two main tasks: Prediction and Description[27,28]. Prediction is an attempt to estimate the possible values of data. The description is an attempt of discovering the interesting patterns in the data. Data mining consists of some key properties such as the Discovery of Patterns, Classification, Prediction of Outcomes, Actionable information and Focusing on large databases.

1. **Automatic Discovery of Patterns:** At this stage, different models are framed to manage data mining. A model is simply defined by using an algorithm to act on a set of data to achieve some patterns automatically. These models are used to mine the data from the framed model.
2. **Classification:** This is a data mining function that designates the collection of items to a target class. This classification predicts the class labels in the target class accurately by classifying a model based on Training set. The class labels of the attribute are classified by defining new data and are tested in a set of test data by comparing the values of known target class.
3. **Prediction:** Prediction is a data mining defines about the identity of one object based on the description of another object. This is quite different from classification. The target class is continuos in prediction stage where as in classification the target class values are discreet. The instances for prediction are time series and regression.
4. **Grouping:** This property in data mining defines about the implementation of groups from the given data.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 16780 - 16790
Received 20 May 2021; Accepted 3 June 2021.

5. **Actionable Information:** Data mining can evolve information by taking some decisions and actions.

Pre-processing or cleaning of data is performed to remove noisy data in data mining. To remove irrelevant attributes from dataset, we performed attribute relevance analysis for prediction analysis. Further, to predict the attribute values regression analysis is conducted by utilizing various regression approaches. Firstly, a dataset with final target values is collected and experimented by examining the specific attribute value in regular intervals of time.

This paper mainly focused the classification of data. Hyper pipes are a simple classifier used for faster classification which is suitable to plot using a large number of attributes. This hyper pipe is constructed for each class in the dataset with each and every point of that class. The samples in this class are classified based on the most similar sample in that particular class. The goal of this work is to determine the performance of the hyperpipes algorithm for different larger data sets[29,30]. The following Section 2 presents the related work. Section 3 explains the methodology with the collection of different datasets. Section 4 describes about the experimental setup and the performance of the classifier among various benchmark datasets. Finally, a conclusion is presented.

II Review of literature

Nuno Laranjeiro, Rui Oliveria, Marco Vieira [1] has explored about effective handling of software bugs by testing web services. In this process vast amounts of response services is classified manually to differentiate from regular services. Apart from consuming large amount of time and effort, it even requires huge manpower to handle such large laborious work of handling errors effectively. The present model is applied on five classification models of text to examine the strength of web services. The effectiveness is measured by using Support Vector Machines, Naïve Bayes, K nearest Neighbor, Hyperpipes, and Large Linear Classification. The evaluation concluded that Hyperpipes has performed well when compared to above mentioned algorithms. The time complexity is also varied with each model; Hyperpipes took 15 minutes to perform execution and is best compared to other defined models.

Zainab Abu Deeb, Thomas Devine and Zongyu Geng [2] have explored in an article “Randomized Decimation HyperPipes” about working of HyperPipes on sparse data sets. The hypothesis is tested by a tool named as Randomized Decimation HyperPipes which allows user to conveniently sea level of sparseness of training data sets[31,32].In this model cell deletion, feature subset selection are introduced.RDH worked better with varying number of selected features on different trials. Hyperpipes did not exhibit better on performance on leaner. So RDH has used its feature selection with combination of deletion in order to increase the datasets sparseness, which performed better.

Authors [3] proposed a method to analyze disasters caused by natural calamities. They occur because of heavy floods, precipitation, river runoffs. Global Warming is the main reason for the occurrence of flood which leads to a great loss of human and infrastructure. The sensors are placed at different areas to gather water fluctuations. The IT technology has not emerged in this field of the issue in a broader way. The present model describes of alert generating system to detect flood and to monitor water level. It provides a notification alerts that describes about the upcoming situation based on determining the water level with sensor technology. The risk and normal situations were analyzed accurately by Random Forest classification model which has generated to an accuracy of 99.5%.

Lars Kotthoff, LAN Miguel and Peter Nightingale has intended in an article [4] “Ensemble Classification for Constraint solver configuration” about instantly selecting algorithms and

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 16780 - 16790
Received 20 May 2021; Accepted 3 June 2021.

automatic parameter tuning. Machine learning classifiers are used to give characteristics of a particular issue. A lot of survey is conducted in algorithm selection and tuning, no similar method of evaluation is taken. The experimental results show that machine learning algorithm performance is variable functioning on new data cannot be made without proper effort. The advantage of this ensemble classification is individual algorithms are combined without having intrinsic knowledge about them. This enables practitioners to apply algorithms without having complete knowledge of them.

Authors [5] have proposed a model "Machine Learning algorithms: "A study on noise sensitivity". The survey was conducted on large data sets where several machine learning algorithms are applied on them and results are tested against noise. The noise is tested on 0 to 0.5 ranges. The dependent variable is converted into nominal from numerical to test classification models. It has described that linear regression has adapted better for gradually increasing noise levels. Decision Table has seemed to be less sensitive to additional noise.

Authors [17] has intended to propose a model which explains about detection of a face used for authentication and verification of person. The fields like Video Surveillance, Human Interaction and Authentication Security are dependent on securing of face detection. This research is performed by gathering 400 faces from Muzaffarabad school students. Apart from this 50 non faces were also collected. Both of them are analyzed and noise reduction techniques and preprocessing are performed. The features are extracted after preprocessing which include geometric features which include image projections, Image cropping etc. The machine learning techniques like Naïve Bayes, Lazy, Misc, Meta, Trees and Rules are used classify non faces and faces. Of all the models used, Hyperpipes has generated with highest accuracy.

Authors [19] proposed a model to analyze the information about trends of E-learning. The computational techniques have shown by E-Learning in the semantic web technology. It focuses on personalized and semantic aware learning than content focused learning pattern. The present model has worked out to analyze the performance of HyperPipes and Naïve Bayes in Indian News context to minimize false positive rates and maximize true positives. The main focus of this model is to make the information available to learner with the framework designed by author. The survey has concluded that HyperPipes performed well in categorizing News.

Authors [20] has intended to propose a model for Intrusion Detection and Prevention about the importance of network security. Everyone needs their communication to be performed in a secured manner. Hacking is the process adopted by intruders to enter into the communication track and disrupts the authenticated flow of control. The cross validation of 3 folds is adopted in this model and classification rate, false negative rate is used to analyze performance. The performance is analyzed and the models like SMO, PART HyperPipes , Naïve Bayes, Random Forest, KStar, Updateable are performed with highest classification rate[33,34,35,36].

III Methodology

HYPERPIPES: Hyperpipes is a simple classifier used for faster classification. This algorithm easily handles huge amount of attributes. A Single pipe is created for each class that keeps the values of attributes in a path without assigning any counts in the training. While in testing, each instance in the pipe is classified with the values and defines the pattern of values of an instance that matches mostly to the pipe. In the Training Phase, a pipe is constructed for each and every class by its label. Once the pipe is created, dataset is scanned by each instance and if the value of the instance is known then the attribute value is added to the pipe. If the value falls within the boundary of pipe, then the attribute value considered may be seen. If it falls outside, then the boundary is updated either by including a new minimum value or maximum value. In Testing Phase, the instances are

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 16780 - 16790
Received 20 May 2021; Accepted 3 June 2021.

compared to each class in the pipes. To decide about an instance in a pipe, the value between the pipe and instance is incremented by a counter. Then the instance of a class is assigned to the pipe and the highest number of matches is shared among the pipe. If there are same matches in multiple pipes, then the highest number of matches from the last pipe can be assigned to the instance of a class. Finally, this hyperpipes do not handle the data sets with numerical attributes or missing values. The main advantage of this algorithm is, it is much faster than any other algorithm and can handle very large quantities of attributes. The following figure depicts the architecture of the proposed method.

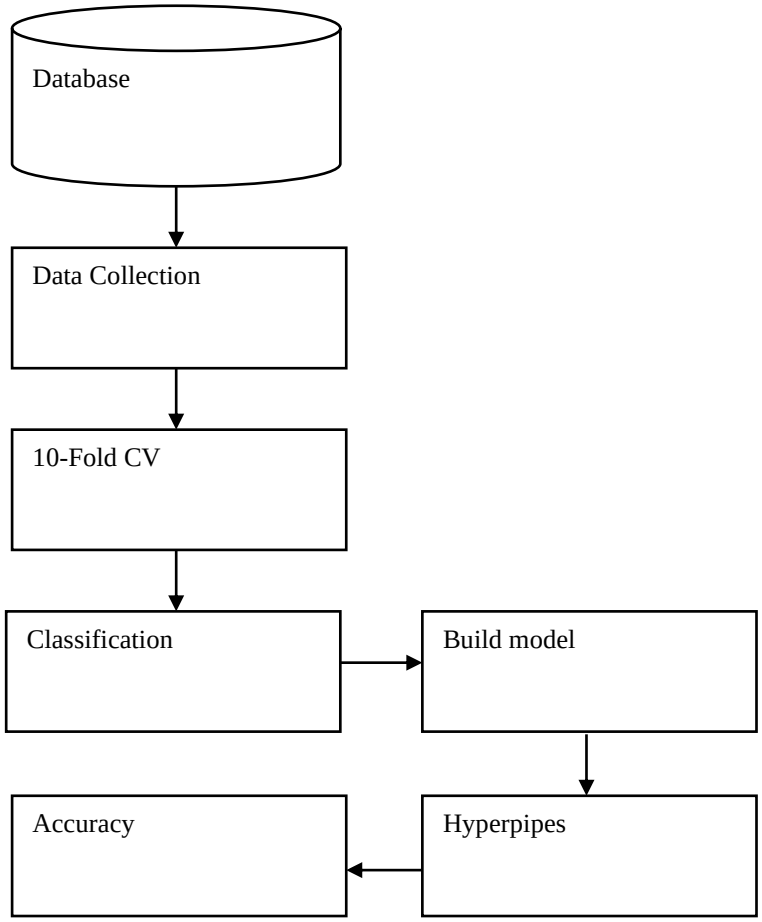

Fig.2. Architecture of Proposed Method.

Data Collection: In this Study, the original datasets have been collected from publicly available UCI repository web site. Based on the collected data, KDDCup, Shuttle, Pendigits, Amazon and Dexter are the original datasets having a large amount of data. KDDCup dataset contains of 35000 samples with 231 attributes, Shuttle dataset contains of 45000 samples with 10 attributes, Pendigits dataset consists of 10922 instances with 17 attributes, Amazon dataset that contains of 1050 samples with 10001 attributes and Dexter dataset consists of 420 instances with 20001 attributes. The following Table lists the number of datasets with instances, attributes and their build time. It also represents the Mean and the Standard Deviations of different data sets.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 16780 - 16790
Received 20 May 2021; Accepted 3 June 2021.

IV Experimental Results

The performance of the Hyperpipes classifier is assessed by collecting the larger datasets. The data sets used in this analysis are collected from UCI database. The database consists of KDD, Shuttle, Amazon, Pendigits, and Dexter. The testing with these datasets exhibits the superior performance with 10-fold Cross Validation.

Table. 1 Statistical details of the datasets.

Dataset	**Instances**	**Attributes**	**Mean**	**Std. Dev**	**Build Time**
KDD	35000	231	10.26	27.77	0.14s
Shuttle	45000	10	48.2	12.25	0.03s
Pendigits	10992	17	38.814	34.258	0.01s
Amazon	1050	10001	11.92	5.69	0.15s
Dexter	420	20001	0.17	3.51	1.03s

With 10-fold cross validation on these datasets, it produces 10 equivalent sized sets by separating each set into Testing and Training in the ratio of 90:10. Then the proposed algorithm is pertained and generates a data set 1. Likewise, it generates for set 2 equal to set 10. Lastly, the performance of the classifiers generates the equal data sets. Finally, the efficiency of the proposed method is also evaluated by evaluation Metrics. True Positive and True Negative shows the positively classified instance and negatively classified instances. False Positive and False Negative shows the symbols and positive instances. These are known as Evaluation Metrics utilized to assess the efficiency of the model in the terms of Accuracy. The Accuracy is evaluated as

$$AC = \frac{TP+TN}{(TP+FP)+(FN+TN)} \quad (1)$$

Correctly Classified Positive Instances are evaluated as

$$TP = \frac{TN}{(FN+TN)} \quad (2)$$

Incorrectly Classified Negative Instances are evaluated as

$$FP = \frac{TN}{(TP)+(FP)} \quad (3)$$

Correctly Classified True Negatives are evaluated as

$$TN = \frac{TP}{(TP+FP)} \quad (4)$$

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 16780 - 16790
Received 20 May 2021; Accepted 3 June 2021.

Incorrectly Classified Positive Instances are evaluated as

$$FN = \frac{FN}{(FN+TN)} \qquad (5)$$

The Following Table shows the Number of Classified and Misclassified Instances.

Table.2. Representation of Classified and Misclassified Instances of the Datasets.

Dataset	**Instances**	**Classified**	**Misclassified**
KDD	35000	34323	677
Shuttle	45000	36607	6893
Pendigits	10992	1527	9465
Amazon	1050	412	638
Dexter	420	344	76

The following Graph represents the accuracy of different UCI datasets.

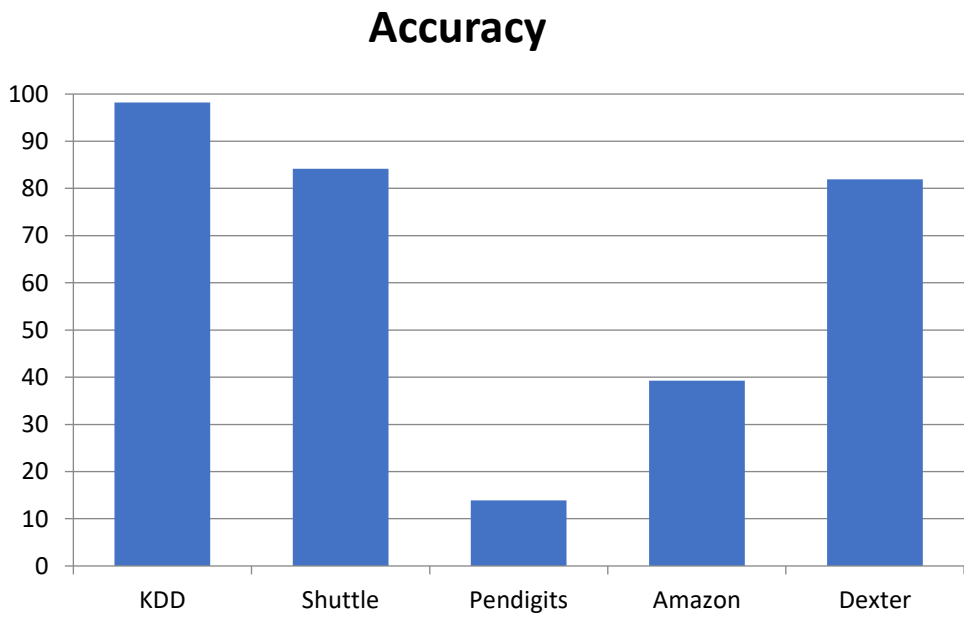

Fig.3. Accuracy Comparison of Various UCI datasets with proposed model.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 16780 - 16790
Received 20 May 2021; Accepted 3 June 2021.

The following Table represents the accuracy rates by class.

Table.3. Class accuracy comparison of various UCI datasets with proposed model

Dataset	**TP**	**FP**	**P**	**R**	**F**	**ROC**
KDDCup	0.981	0.979	0.965	0.981	0.973	0.573
Shuttle	0.842	0.575	0.816	0.842	0.77	0.827
Pendigits	0.139	0.100	0.369	0.066	0.066	0.885
Amazon	0.392	0.014	0.663	0.392	0.408	0.94
Dexter	0.819	0.175	0.839	0.819	0.817	0.933

V Conclusion

In this work, we adopted Hyperpipes is the proposed method which is a simple classifier used for faster classification. This hyperpipes do not handle the data sets with numerical attributes or missing values. The main advantage of this algorithm is, it is much faster than any other algorithm and can handle very large quantities of attributes. For this experiment, the efficiency was evaluated on various large sized data sets that contain more than 10000 instances or attributes. Thereby we founded that under these five datasets, KDDCup dataset with 35000 instances and 231 attributes which is one of the benchmark dataset generates the highest accuracy of 98.20% when compared with other datasets.

References

[1] Nuno Laranjeiro, Rui Oliveria, Marco Vieira," Classification Algorithms in Web Services Robustness Testing", " 29 th IEEE International Symposium on Reliable Distributed Systems".

[2] Zainab Abu Deeb, Thomas Devine and Zongyu Geng , "Randomized Decimation HyperPipes", West Virginia University, Morgantown, VVV.

[3] Mohammed Khalaf, Abir Jaafar Hussain, Dhiya Al-Jumeily, Paul Fergus, Ibrahim Olatunji Idowu," Advance Flood Detection and Notification System based on Sensor Technology and Machine Learning Algorithm".

[4] Lars Kotthoff, Lan Miguel and Peter Nightingale , "Ensemble Classification for Constraint solver configuration", University of St Andrews.

[5] Elias Kalapanidas, Nikolaos Avouris_Marian Craciun and Daniel Neagu," Machine Learning algorithms : a study on noise sensitivity".

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 16780 - 16790
Received 20 May 2021; Accepted 3 June 2021.

[6] Fco.Javier Ord o nez , Agapito Ledezma and Arceli Sanchis,"Genetic Approach for Optimizing Ensembles of Classifiers",21st International FLAIRS Conference,2008.

[7] Roxana Danger,Isabel Segura-Bedmar,Paloma Martinez,Paolo Rosso, "A comparison for machine learning techniques for detection of drug target articles", "Journal of Biomedical informatics 43,2010.

[8] Agapito Ledezma,Ricardo Aler,Araceli Sanchis and Daniel Borrajo,"Empirical Evaluation of Optimized Stacking Configuratons".

[9] Kyung Choi,Xinyi Chen, Shi Li,Mihui Kim ,Kijoon Chae and JungChan Na, "Intrusion Detection on NSM based Dos Attacks Using Data Mining in Smart Grid, Energies ,October 2012.

[10] Animut Belay Asres, "A Semi Supervised Approach for Amharic News Classification" ,June 2012.

[11] Eya Ben Ahmed,Ahlem Nabliz and Falez Gorgouri,"Building Multiview Analyst Profile From Multidimensional Query Logs:From Consensual to Conflicting Prefernces","IJCSI International Journal for Computer Science Issues" Vol.9,January 2012.

[12] Radim Rehurek, Petr Sojka, "Automatic Classification and Categorization of Mathematical Knowledge".

[13] I.M. Scott, W. Lin, M. Liakata, J.E. Wood, C.P. Vermeer, D. Allaway, J.L. Ward,J. Draper, M.H. Beale, D.I. Corol, J.M. Baker, R.D. King,"Merits of Random Forests emerge in evaluation of chemometric classifiers by external validations", "Analytica Chimica Acta", 2013.

[14] Yan Ma,Lan Guo, Bojan Cukic, "A Statistical Framework for the prediction of Fault-Proneness".

[15] Theo_lis George-Nektarios, "Weka Classifiers Summary", Athens University of Economics and Business ,November21, 2013.

[16] Manuel Fern_andez-Delgado,Eva Cernadas,Sen_en Barro,"Do we need hundreds of classifiers to solve real world classification problems?","Journal of Machine Learning Research 15",2014.

[17] Lal Huussain," Classification of Human Faces and Non Faces Using Machine Learning Techniques, "International Journal of Electronics and Electrical Engineering",Vol.2,No.2,June,2014.

[18] Urmila Mahor,Sujoy Das, "Performance Evaluation of Various Feature Extraction and Classification Techniques for Authorship Attribution", "International Journal of Innovation and Scientific Research", Vol.16,No.1,June 2015.

[19] DR.Sushil Kumar Rameshpant Kalmegh, "International Journal of Pure and Applied Research in Engineering and Technology", IJPRET, Vol 4, January 5, 2016.

[20] Mr.A.Siles Balasingh, Mr.N.Surenth ,"An approach of Data Mining Algorithm for Intrusion Detection and Prevention System", IOSR Journal of Computer Engineering,Vol.6,October 2016.

[21] Lokaiah Pullagura, Madhusudhan Rao Dontha, Srinivasarao Kakumanu, "Recognition of Fetal Heart Diseases Through Machine Learning Techniques", Annals of the Romanian Society for Cell Biology,2021.

[22] Lokaiah Pullagura, Dr. Anilkumar, "Analysis of Train Delay Prediction System based on Hybrid Model", Journal Of Advanced Research in Dynamical & Control Systems, Vol. 12, 07-Special Issue, 2020.

[23] Lokaiah Pullagura, Dr. Anilkumar, "An Effective LSTM Network Model For Accurate Prediction of Delays in Indian Railway Networks", International Journal of Advanced Trends in Computer Science and Engineering, Vol 9, No.4, July-August 2020.

[24] P. Lokaiah, S. Nyamathulla, M. Kiran Kumar, KBV Rama Narasimhamam, "Performance Evaluation of Different Machine Learning Techniques for Prediction of Diabetes, Journal of Critical Reviews, doi: 10.31838/jcr.07.18.50.

[25] Dr. P. Ratnababu, P. Lokaiah, "An Effective noise reduction technique for class imbalance classification", International Journal of Psychosocial Rehabilitation, Vol 24, Issue 04, 2020, ISSN 1475-7192.

[26] Bhowmik, C., Ghantasala, G. P., & AnuRadha, R. (2021). A Comparison of Various Data Mining Algorithms to Distinguish Mammogram Calcification Using Computer-Aided Testing Tools. In *Proceedings of the Second International Conference on Information Management and Machine Intelligence* (pp. 537-546). Springer, Singapore.

[27] Patan, R., Ghantasala, G. P., Sekaran, R., Gupta, D., & Ramachandran, M. (2020). Smart healthcare and quality of service in IoT using grey filter convolutional based cyber physical system. *Sustainable Cities and Society*, *59*, 102141.

[28] Krishna, N. M., Sekaran, K., Vamsi, A. V. N., Ghantasala, G. P., Chandana, P., Kadry, S., ... & Damaševičius, R. (2019). An efficient mixture model approach in brain-machine interface systems for extracting the psychological status of mentally impaired persons using EEG signals. *IEEE Access*, *7*, 77905-77914.

[29] Chandana, P., Ghantasala, G. P., Jeny, J. R. V., Sekaran, K., Deepika, N., Nam, Y., & Kadry, S. (2020). An effective identification of crop diseases using faster region based convolutional neural network and expert systems. *International Journal of Electrical and Computer Engineering (IJECE)*, *10*(6), 6531-6540.

[30] Ghantasala, G. P., & Kumari, N. V. (2021). Identification of Normal and Abnormal Mammographic Images Using Deep Neural Network. *Asian Journal For Convergence In Technology (AJCT)*, *7*(1), 71-74.

[31] Ghantasala, G. P., Kallam, S., Kumari, N. V., & Patan, R. (2020, March). Texture Recognization and Image Smoothing for Microcalcification and Mass Detection in Abnormal Region. In *2020 International Conference on Computer Science, Engineering and Applications (ICCSEA)* (pp. 1-6). IEEE.

[32] Ghantasala, G. P., Tanuja, B., Teja, G. S., & Abhilash, A. S. (2020). Feature Extraction and Evaluation of Colon Cancer using PCA, LDA and Gene Expression. *Forest*, *10*(98), 99.

[33] Kumari, N. V., & Ghantasala, G. P. (2020). Support Vector Machine Based Supervised Machine Learning Algorithm for Finding ROC and LDA Region. *Journal of Operating Systems Development & Trends*, *7*(1), 26-33.

[34] Ghantasala, G. P., Reddy, A., Peyyala, S., & Rao, D. N. (2021). Breast Cancer Prediction In Virtue Of Big Data Analytics. *INTERNATIONAL JOURNAL OF EDUCATION, SOCIAL SCIENCES AND LINGUISTICS*, *1*(1), 130-136.

[35] Ghantasala, G. P., Reddy, A. R., & Arvindhan, M. Prediction of Coronavirus (COVID-19) Disease Health Monitoring with Clinical Support System and Its Objectives. In *Machine Learning and Analytics in Healthcare Systems* (pp. 237-260). CRC Press.

[36] Sreehari, E., & Ghantasala, P. G. (2019). Climate Changes Prediction Using Simple Linear Regression. *Journal of Computational and Theoretical Nanoscience*, *16*(2), 655-658.