IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2320-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 10, Issue 2, February 2021

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.512

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 10, Issue 2, February 2021 ||

DOI:10.15680/IJIRSET.2021.1002016

Machine Learning Techniques and Algorithms: A Survey

Renuka R. Patil[1], Bharathi S. Shetty[2], Suresha[3]

Faculty of Dept. of Computer Science & Engineering, Sri Venkateshwara College of Engineering, Bangalore, Karnataka, India [1]

Faculty of Dept. of IT, WCE. Sangli, Maharashtra, India[2]

Principal, Sri Venkateshwara College of Engineering, Bangalore, Karnataka, India [3]

ABSTRACT: In the current era of emerging technologies, data is generated in huge quantity from each and every domain of the technologies. These data play a vital role in predicting the future. To predict the future using these data is a challenging task of Machine Learning (ML). The data collected is not only huge in quantity but also in size. Hence, manual prediction of the future is a complex and tedious task. To overcome with this problem, the machine is to be trained to predict the future with the help of this collected data. The data is classified into train data set and test data test. To train the machine, there are various ML techniques and algorithms are available. There are various algorithms in Machine Learning algorithms to train the model. This paper mainly focuses on the ML techniques and algorithms, importance of machine learning in data science and their applications in various fields.

KEYWORDS: Challenging Task, Machine Learning, Manual Prediction, Train Data Set, Test Data Set

I. INTRODUCTION

Machine learning is one of the applications and is a subset of Artificial Intelligence (AI) [2]. ML is an ability of machine to learn automatically and improve the future performance based on the past experience. ML technique functions without being programmed explicitly. ML mainly focuses on developing the computer programs which access the data and perform further operations. The Figure 1 briefs the relation between AI, ML and Deep Learning (DL).

Machine Learning basically functions on modelling which uses ML techniques and algorithms, in which the machine studies from the past data similar to humans who learn from the past experiences [1].

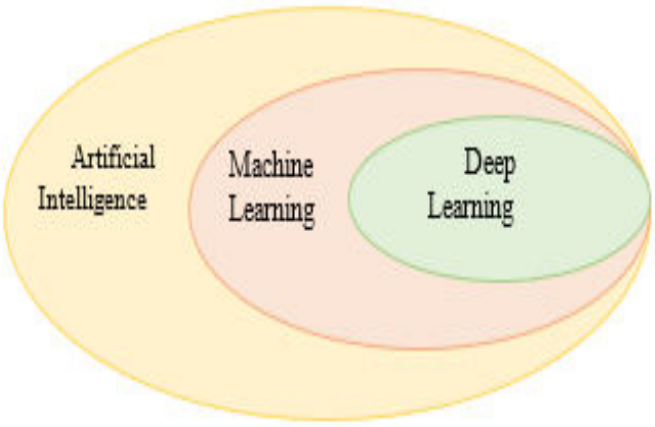

Figure 1: Relation between AI, ML & DL

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 10, Issue 2, February 2021 ||

DOI:10.15680/IJIRSET.2021.1002016

ML is mainly used in the industries. Machine Learning algorithms are classified as Supervised Learning (is also called as a Linear Regression), Unsupervised Learning and
Reinforcement Learning algorithms. Supervised learning algorithms represent the labels and functions based on the previous data. Unsupervised learning algorithm does not represent the labels. ML models are classified into the following three types of models based on the task performed:

1. Regression
2. Classification
3. Clustering

From the above ML models, Regression and Classification fall under the Supervised Learning Algorithms and Clustering is an Unsupervised Learning Algorithm [5].
This paper is organized as; the second section briefs about the literature survey about the Machine Learning algorithms and techniques, section 3 Machine Learning-Regression Algorithms, section 5 focuses on the role of ML in Datascience and section 6 ends with the conclusion about this paper.

II. LITERATURE SURVEY

This section briefs the research carried out in the field of Machine Learning and its applications in various fields.
Fugui Liu et al. [14] applied Support Vector Machine (SVM) with Radial Basis Function (RBF) kernel function for diagnosis of the Heart disease. These algorithms are implemented based on the statistical learning theory. From the dataset, to select the suitable parameters for Kernel function Grid search technique is applied which also improves the accuracy of elevated classifications. The survey is carried out to analyze the suitable ML algorithm for the diagnosis of Heart disease.
Mahdi Motavalli Haghighi et al. [15] applied Fuzzy Neural Networks along with the adoption of SVM algorithm and Artificial Neural Networks (ANN) to identify and diagnose the hepatitis and thyroid diseases. The researchers continued the research to identify the varieties and stages of an ailment that incorporates 6 classes for the hepatitis ailment-B. The survey is carried out to find the various stages of ailment that integrate the six-classes of Hepatitis ailment (2-stages of Hep-B and Hep-C), non-Hepatitis and five-classes of Thyroid diseases (Thyrotoxicosis, Hypothyroid, Subclinical thyrotoxicosis, subclinical hypothyroid and non-thyroid disease). It is also concluded that, for Hepatitis illness, an excellent accuracy of 98% and for Thyroid ailment of 99%. The Table 1 briefs about the results of survey carried out in various algorithms ML in different areas.

Table 1: Results of the survey carried out considering various ML algorithms in different areas

Sl. No.	Authors	Methodology (Algorithm)	Dataset
1	Tomar et al.[18]	Support Vector Machine (SVM)	Heart disease
2	Ghosh S. et al.[19]	SVM & MLP-BPN	Breast Cancer
3	Basu et al.[20]	Linear Kernel with SVM	Erythema-to-Squamous Illness
4	Seyede et al.[21]	GA-SVM	DNA microarray disease
5	Amine LazouniMohammed E et al.[22]	Various Kernel Functions with SVM	Anesthetic Examination
6	Tipawan et al.[23]	SVM ensembles with Correlation- based selection	Cardiotocography
7	Austeclino et al.[24]	Bayesian Network and ANNs	Malaria Asymptomatic disease
8	Luis et al.[26]	WFE and SVM	Alzheimer's disease
9	Carlos et al.[25]	ANNs	Brain Tumor
10	Mrudula et al.[27]	ANNs and SVM	Heart Illness

In the next section, the different ML algorithms are surveyed.

III. MACHINE LEARNING-REGRESSION ALGORITHMS

In section 1, the three different models such as Regression, Classification and Clustering. In this section the algorithm. Here, the attention has been given more on the prediction of future results by using linear regression concepts. Broadly

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 10, Issue 2, February 2021 ||

DOI:10.15680/IJIRSET.2021.1002016

speaking, it is a form of predictive modelling technique which tells about the relationship between the dependent (target variable) and independent variables (predictors). The two types of linear regression are:

- Simple linear regression
- Multiple linear regression

Simple Linear Regression:

The most elementary type of regression model is the simple linear regression which explains the relationship between a dependent variable and one independent variable using a straight line. The straight line is plotted on the scatter plot of these two points. The strength of the linear regression model can be assessed using 2 metrics:

- R2 or Coefficient of Determination
- Residual Standard Error (RSE)

R2 or Coefficient of Determination:

This is also an alternative way of checking the accuracy of the model, which is R2 statistics. In general term, it provides a measure of how well actual outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model, i.e. expected outcomes. Overall, the higher the R-squared, the better the model fits the data.

Mathematically, it is represented as: R2 = 1 - (RSS / TSS)

Here, RSS (Residual Sum of Squares) is defined as the total sum of error across the whole sample. It is the difference between the actual and expected output. And, TSS (Total sum of squares) is the sum of errors of the data points from mean of response variable.

Bayes Theorem:

The probability of an event is the measure of the chance that the event will occur as a result of an experiment. The probability of an event A is the number of ways event A can occur divided by the total number of possible outcomes.

Probability is the chance of occurrence of an event, it can be defined as a ratio of the number of desired outcomes to the number of total possible outcomes. It is denoted as p(E), indicating the probability of an event E. In the numerator we have the number of favorable outcomes for this event E and in the denominator, we have the total number of outcomes corresponding to our data.

Logistic Regression:

A binary classification is basically, it is a classification problem in which the target variable has only 2 possible values, or in other words, two classes. Some examples of binary classification are –

1. A bank wants to predict, based on some variables, whether a particular customer will default on a loan or not.
2. A factory manager wants to predict, based on some variables, whether a particular machine will break down in the next month or not.
3. Google's backend wants to predict, based on some variables, whether an incoming email is spam or not.

Sample Selection for Logistic Regression:

Selecting the right sample is essential for solving any business problem. There are major errors and are noted as;

- Cyclical or seasonal fluctuations in the business that need to be taken care of while building the samples. E.g. Diwali sales, economic ups and downs etc.
- The sample should be representative of the population on which the model will be applied in the future.
- For rare events samples, the sample should be balanced before they are used for modelling.

Segmentation:

It is very helpful to perform segmentation of population before building a model.

WOE Transformation

Three are three important things about WOE: • Calculating woe values for fine binning and coarse binning • Importance of woe for fine binning and coarse binning • Usage of woe transformation There are two main advantages of WOE: WOE reflects group identity, this means it captures the general trend of distribution of good and bad customers. E.g. the difference between customers with 30% credit card utilization and 45% credit card utilization is not

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 10, Issue 2, February 2021 ||

DOI:10.15680/IJIRSET.2021.1002016

the same as the difference between customers with 45% credit card utilization and customers with 60% credit card utilization. This is captured by transforming the variable credit card utilization using WOE. Secondly, WOE helps you in treating missing values logically for both types of variables - categorical and continuous. E.g. in the credit card case, if you replace the continuous variable credit card utilization with WOE values, you would replace all categories mentioned above (0%-45%, 45% - 60%, etc.) with certain specific values, and that would include the category "missing" as well, which would also be replaced with a WOE value.

These are the important Machine Learning Algorithms and the next section deals about the Role of Machine Learning in Datascience.

IV. ROLE OF ML IN DATASCIENCE

Data analytics is typically aimed towards solving a problem to generate insights from data. Typically, the process of analysis has to go through the following steps to obtain a solution for a data analytics problem:

- Start by developing a business understanding around the problem. This is the first step because understanding the problem and its impact on the business is very crucial.
- After developing the business understanding, it is required to understand the various data sets or sources of data which can be leveraged to solve the problem at hand. This is the second step in the CRISP DM framework and is called data understanding.
- The third step is called data preparation. It is the most important and time-consuming step in the entire analysis. Once the data sets have been figured and understood, they need to be prepared in various ways for the analysis to happen.
- The fourth step – modelling – is the most interesting step of the entire CRISP DM process. After preparing the data, models are built on it to solve the business problem at hand and generate insights from the data.
- Before the model is deployed, it needs to be evaluated to check its accuracy, usefulness and to understand how well it is performing. Model evaluation is a continuous step and needs to be performed on the final model for the model to be valid over time.
- The final step in the framework is model deployment. Once the model passes the evaluation criterion, it is ready for deployment.

Business Understanding:

The first stage of the framework is to develop a business understanding. For this, two steps are required to carry out. They are:

- Determine the business objective
- Identify the goal of the data analysis

Determining the business objective is of utmost importance. Until the business objectives have been finalised, the data cannot be collected or worked upon. Then, it is required to determine the goals of data analysis, and work towards achieving them in the CRISP-DM framework.

Data Understanding:

This stage comprises of four key steps to understand the available data, and identify new relevant data in order to solve the business problem.

- **Collect relevant data-** First, it is necessary to identify and collect the right set of data sets that can be used for the analysis. They can be available within the firm, or to collect the data from other sources such as open source repositories or government data sets.
- **Describe data**- for explicit information: Once the data set is identified,it is required to describe its contents and explore insights to better understand the data and its business implications. To describe the CrunchBase data, the data dictionary is to be created, that lists down the types of variables (e.g. sectors, company names, etc.), the number of records, and the types of analysis.
- **Explore data-** for implicit insights: To explore data, the different types of graphs are plotted on Excel/R, e.g. to understand the range of funding received by companies.
- **Verify data quality-** to remove errors: After understanding the data structure, the quality of the data is to be examined and it addressesvarious factors. They are;

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 10, Issue 2, February 2021 ||

DOI:10.15680/IJIRSET.2021.1002016

- Is the data complete, does it cover all the cases and records?
- Is the data correct, or does it contain errors and, if there are errors, how common are they?
- Are there missing values in the data? If so, how are they represented?
- Where do the missing values occur, and how common are they?

All the issues with the data are to be understood here so that they can be taken care of in the next steps. For example, in the CrunchBase data set, funding for a certain company is inaccurately reported as negative $-20K or $2Bn, when the normal reported funding ranges between 0 - $20M.

Data Preparation:

Data preparation is the most important and time-consuming step of the CRISP-DM framework and is carried out in the following steps:

- Data which is of interest is selected. It may be spread across different files or sources.
- Then, the data is to be integrated from these multiple sources. For example, in the CrunchBase data, information is spread across different files and we integrate them together to solve a particular business problem.
- After data integration comes the data preparation stage. Missing value treatment, outlier treatment, and removing the erroneous values are a few major components of the data cleaning process.
- Constructing the data is the next step, which involves the creation of new features originally not present in the data set. The purpose of this step is to increase the information we get from the data and to reduce the number of variables originally present in the data set.
- Finally, format the data. In this step, no changes are made to the data in the data set, but to its structure. The variables present in the data set are set to the correct format as required by the analysis tool.

Data Modelling:

Modelling activity in the CRISP-DM framework involves two major tasks. The first task is to understand the problem domain and select the appropriate family of models that is suitable for solving the problem at hand. The second task is to select appropriate algorithms for creating the model from the chosen family of models.

Model Evaluation:

The predictive models can be tested to assess their effectiveness in solving the problem. This is the fifth stage of the framework – model evaluation. Modelling and evaluation together is an iterative process in which the models are tweaked until satisfactory evaluation results are obtained.

Model Deployment:

This is the last stage of the framework, where the model is translated into a business strategy. Business data is fed into the model and the model results are used to inform business decisions on an on-going basis.

The CRISP-DM framework does not end at the last stage of model deployment. The important thing to note is that CRISP-DM is an iterative process. For example, data understanding can enhance the business understanding. Similarly, after model evaluation, if the model does not perform great, it is required to go back to the data preparation stage, and then develop the model again.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 10, Issue 2, February 2021 ||

DOI:10.15680/IJIRSET.2021.1002016

Fig.2, describes the process of CRISP-DM Framework.

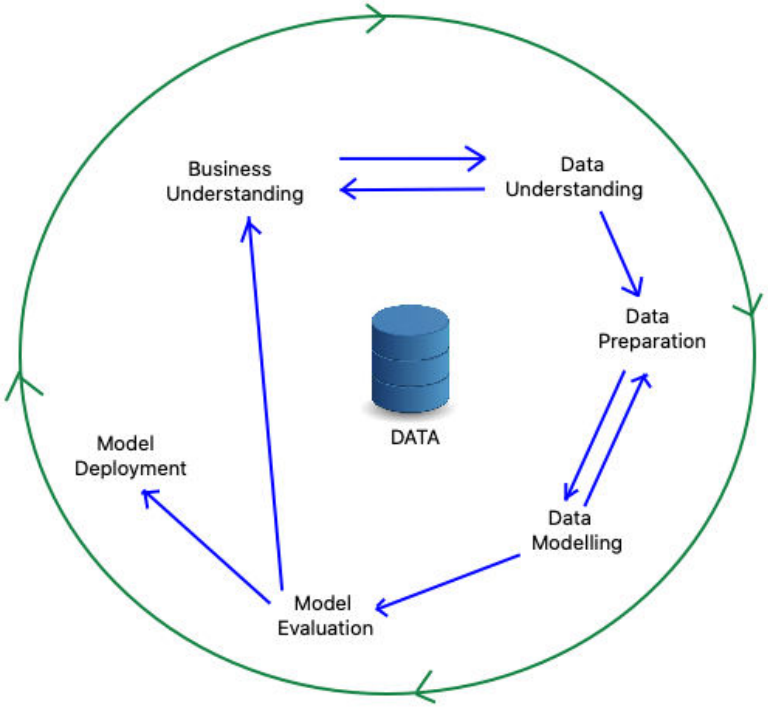

Figure 2:CRISP-DM Framework

Data science refers to the process of extraction of useful insights from data. This interdisciplinary approach merges various fields of computer science, scientific processes and methods, and statistics in order to extract data in automated ways. In order to mine big data, which is closely associated with the field, data science uses a diverse range of techniques, tools and algorithms gleaned from the fields.

Machine Learning is an important sub area of Artificial Intelligence [28]. It builds computers to get into a self-learning mode without any explicit programming. When machine is fed with new data-set, they learn, change and predict the future by themselves. The era of ML has been around for a while now [29]. However, an ability to quickly and automatically apply mathematical calculations over the data science.

Machine learning has been applied in various areas and domains such as, the self-driving Google car, friend recommendations on Facebook, the online recommendation engines, in cyber fraud detection, in offer suggestions from Amazon and many more [30].

V. CONCLUSION

ML is one of the fastest emerging technologies in the field of Technology. In this work, a widespread survey is carried out considering the existing ML algorithms which include the usage of semi-automated and automated medical analysis applied in the application of explicit diseases. The applications of various neural networks, K-Nearest Neighbor (KNN) and Bayesian networks are focused in this research survey.

REFERENCES

[1] L. Hui and D. Pi (2017).Integrative Method Based on Linear Regression for the Prediction of Zinc binding Sites in Proteins. IEEE Access, 5, 14647-14656. doi:10.1109/access.2017.2731872

[2] L. Wang and D. Wang (2017). Intelligent CFAR Detector Based on Support Vector Machine. IEEE Access, 5, 26965-26972.

[3] J. Tanha (2015). Semi-supervised self-training for decision tree classifiers. International Journal of Machine Learning and Cybernetics, 8(1), 355-370. doi:10.1007/s13042-015-0328-7

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 10, Issue 2, February 2021 ||

DOI:10.15680/IJIRSET.2021.1002016

[4] D. Khadim, M. Fleur and D. Gayo (2016). Large scale biomedical texts classification: a k-NN and an ESA-based approaches. Journal of Biomedical Semantics, 7, 40. doi:10.1186/s13326-016-0073-1
[5] R. Hong, H.M. Wang and L. Jian (2016). Privacy-Preserving k-Nearest Neighbor Computation in Multiple Cloud Environments. IEEE Access, 4, 9589-9603.
[6] L. Jiang and C. Li (2016). Deep feature weighting for naive Bayes and its application to text classification. Journal of Engineering Applications of Artificial Intelligence, 52, 26-39. doi:10.1016/j.engappai.2016.02.002
[7] M. Ahmed and H. Alison (2018). Modeling built-up expansion and densification with multinomial logistic regression cellular automata and genetic algorithm. 67, 147-156.
[8] T. Razzaghi and R. Oleg (2016). Multilevel Weighted Support Vector Machine for Classification on Healthcare Data with Missing Values. PLUS ONE, 1-18. doi:10.1371/journal.pone.0155119
[9] L Hui and D. Pi, "Integrative Method Based on Linear Regression for the Prediction of Zinc binding Sites in Proteins", IEEE Access, vol. 5, pp. 14647-14656, August 2017. (Repeated)
[10] L. Wang and D. Wang, "Intelligent CFAR Detector Based on Support Vector Machine", IEEE Access, vol. 5, pp. 26965-26972, December 2017. (Repeated)
[11] R. Enrico and L. Michel (2016). The Counter a Frequency Counter Based on the Linear Regression. IEEE Transactions on Ultrasonics Ferroelectrics, 63(7), 961-969. doi:10.1109/tuffc.2016.2570604
[12] J. Han, M. Kamber and J. Pei (2012). Data Mining: Concepts and Techniques. MK Series.
[13] David and B. Lerner (2004). Pattern Classification Using a Support Vector Machine for Genetic Disease Diagnosis. Proceedings. 23rd IEEE Convention of Electrical and Electronics Engineers in Israel, 289-292. doi:10.1109/eeei.2004.1361148
[14] F. Liu, Y. Zhang, Z. Zhao, D. Li, X. Zhou and J. Wang (2012). Studies on application of Support Vector Machine in diagnose of coronary heart disease. Sixth International Conference on Electromagnetic Field Problems and Applications, 1-4. doi:10.1109/icef.2012.6310380
[15] M.M. Haghighi and S. Shariati (2010). Comparison of ANFIS Neural Network with Several Other ANNs and Support Vector Machine for Diagnosing Hepatitis and Thyroid Diseases. International Conference on Computer Information Systems and Industrial Management Applications, 596-599. doi:10.1109/cisim.2010.5643520
[16] H.I. Elshazly, A.M. Elkorany and A.E.Hassanien (2014). Lymph diseases diagnosis approach based on support vector machines with different kernel functions. 9th International Conference on Computer Engineering & Systems, 198-203. doi:10.1109/icces.2014.7030956
[17] D. Tomar, B.R. Prasad and S. Agarwal (2014). An efficient Parkinson disease diagnosis system based on Least Squares Twin Support Vector Machine and Particle Swarm Optimization. 9th International Conference on Industrial and Information Systems, 1-6. doi:10.1109/iciinfs.2014.7036603
[18] T. Divya and S. Agarwal (2014). Feature Selection based Least Square Twin Support Vector Machine for Diagnosis of Heart Disease. International Journal of Bio-Science & Bio-Technology, 6(2). doi:10.14257/ijbsbt.2014.6.2.07
[19] S. Ghosh, S. Mondal and B. Ghosh (2014). A Comparative Study of Breast Cancer Detection based on SVM and MLP BPN Classifier. First International Conference on Automation,Control, Energy and Systems, 1-4. doi:10.1109/aces.2014.6808002
[20] Basu, S.S. Roy and A. Abraham (2015). A Novel Diagnostic Approach Based on Support Vector Machine with Linear Kernel for classifying the erythematosquamous disease. International Conference on Computing Communication Control andAutomation, 343-347.
[21] S. Z. Paylakhi, S. Ozgoli and S. H. Paylakhi (2015). A novel gene selection method using GA/SVM and Fisher criteria in Alzheimer's disease. IEEE 23rd Iranian Conference on Electrical Engineering. doi:10.1109/iraniancee.2015.7146349
[22] M.E.A. Lazouni and N. Settouti (2014).An SVM Intelligent System for Pre-anesthetic Examination. IEEE Second World Conference on Complex Systems, 73–78. doi:10.1109/icocs.2014.7060908
[23] T. Silwattananusarn, W. Kanarkard and K.Tuamsuk (2016). Enhanced Classification Accuracy for Cardiotocogram Data with Ensemble Feature Selection and Classifier Ensemble. Journal of Computer and Communications, 4, 20-35.
[24] A.M.B. Junior, A.A. Duarte, M.B. Netto and B.B. Andrade (2010). Artificial Neural Networks and Bayesian Networks as Supporting Tools for Diagnosis of Asymptomatic Malaria. 12th IEEE International Conference on e-Health Networking Applications and Services (Healthcom), 106 – 111.
[25] C. Arizmendi, D.A. Sierra, A. Vellido and E. Romero (2011). Brain Tumour Classification Using Gaussian Decomposition and Neural Networks. 33rd Annual International Conference of the IEEE EMBS Boston, Massachusetts USA, 5645-5648. doi:10.1109/iembs.2011.6091366

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 10, Issue 2, February 2021 ||

DOI:10.15680/IJIRSET.2021.1002016

[26] L.J. Herrera, I. Rojas, H. Pomares, A. Guillén, O. Valenzuela and O. Baños (2013). Classification of MRI images for Alzheimer's disease detection.SocialCom/PASSAT/BigData/EconCom/BioMedCom, 846-851. doi:10.1109/socialcom.2013.127

[27] M. Gudadhe, K. Wankhade and S.Dongre. Decision Support System for Heart Disease based on Support Vector Machine and Artificial Neural Network. IEEE International Conference on Computer & Communication Technology (ICCCT'10), 741-745. doi:10.1109/iccct.2010.5640377

[28] S. Saha and S. Ghorai (2015). Effect of Feature Fusion for Discrimination of Cardiac Pathology. Third International Conference on Computer, Communication, Control and Information Technology (C3IT),1-6. doi:10.1109/c3it.2015.7060202

[29] R. Suganya and S.Rajaram (2013). Feature Extraction and Classification of Ultrasound Liver Images using Haralick Texture-Primitive Features: Application of SVM Classifier. International Conference on Recent Trends in Information Technology, 596- 602. doi:10.1109/icrtit.2013.6844269

[30] M. Fathurachman, U. Kalsum, N. Safitri and C.P. Utomo (2014). Heart Disease Diagnosis using Extreme Learning Based Neural Networks. International Conferencse of Advanced Informatics: Concept, Theory and Application, 23-27. doi:10.1109/icaicta.2014.7005909

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor:
7.512

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर
NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details