A Clinical Perspective of Comparison on Prediction Models for Heart Failure Risk

Divya G
Department of Computer Science & Engineering, VFSTR Deemed to be University
Vadlamudi, Guntur, Andhra Pradesh, INDIA

Hemantha Kumar Kalluri
Department of Computer Science & Engineering, VFSTR Deemed to be University
Vadlamudi, Guntur, Andhra Pradesh, INDIA

V Ramakrishna Sajja
Department of Computer Science & Engineering, VFSTR Deemed to be University
Vadlamudi, Guntur, Andhra Pradesh, INDIA

DS Bhupal Naik
Department of Computer Science & Engineering, VFSTR Deemed to be University
Vadlamudi, Guntur, Andhra Pradesh, INDIA

Abstract- India is one of the preeminent region challenging downfalls because of heart disease. The utmost essential task in medical organizations is to avoid infections. Many people can be saved if the disease is predicted quickly. Machine learning classification methods can undoubtedly benefit the medical field by providing proper and immediate prevention of illness. The objective is to correctly and accurately classify records into the presence or absence of Heart disease. The experiments were conducted on Cleveland, Hungarian, VA long beach, and Switzerland from UCI repository using Decision Tree, Support Vector Machine, Random Forest Classifier, K Nearest Neighbor, Convolution Neural Networks, Logistic Regression algorithms.

Keywords – Heart disease, classification, Heart failure risk, Convolution Neural Networks

I. INTRODUCTION

Heart Failure, also indicated as Congestive Cardiac Failure(CCF) appears when heart muscles come elegant and are not able to supply a sufficient quantity of hemoglobin to the heart. Heart Failing is a complicated and responsive effort. Prognosis of heart disease is considered as one of the utmost significant subjects in the field of the scientific information study. The quantity of data in the medical organization is massive. Data mining turns a vast number of unprocessed health care data into knowledge that can use to form knowledgeable outcomes and predictions.

Heart Failure Risk Factors:

1) **Age:** The maximum of heart failures appear in people ages 65 and earlier, but male risks begin to rise at 45(for female, it seems at 55)
2) **Family History and Genetics:** persons with a family history of cardiomyopathy (a disease that affects the heart muscle) are at high risk of heart failure
3) **Emotions:** chronic stress and anxiety can take a tool. They may take arteries to tighten, increasing blood pressure. Less Psychological health might also prompt inactivity and a bad diet.
4) **Tobacco Use:** Fume can harm the heart and blood vessels, which increases heart failure
5) **Obesity**: Overweight puts significant stress on the heart and heart attack risk will increase if the patent has also had hypertension and high cholesterin.

A recent study depleted in 2019 by world health organization represents the outcome that 56.9 total deaths appear in the earth all along the duration 2016 is owed to heart disease**.** Heart failure is the most type of coronary disease, killing around 370,000 persons each year. World Health Organization (WHO) recognized the potential of data mining that it can provide to predict the initial phase of coronary disease and can contribute to the exact result of the disease. Data mining is mostly the analysis of information from a massive number of unprocessed data. It is also known as sub-part of data management.

A.Motive

The primary incitation of this analysis is to give a correct disease diagnosis framework with a shortened feature set. Usually, doctors have to attain more number of tests to analyze an appropriate disease which desires the amount of capital and space. The automated disease prognosis system will predict coronary syndrome with the exact outcome in the area and work reduction.

B.Research philanthropy

Following are the few contributions of the expected analysis

- Developing previous standard systems
- Prognosis of coronary disease
- Increase Accuracy
- Developing performance
- Developing validity

II. LITERATURE REVIEW

Tiwaskar et al. [1] have proposed character-level convolution neural networks for large scale text classification. The achievement of statistical evaluation, Decision tree classifier, Random Forest classifier, and CNN are compared on a dataset collected from UCI Repository (Cleveland). They examined various efficiency parts, and according to their study, Convolution neural network is giving excellent outcomes as related to other methods.

Liaqat et al. [2] have proposed an automated diagnostic $\chi 2$ statistical model for features refinement and DNN for classification. The performance of the method was compared with other well-known methods like Adaboost, Decision Tree, Convolution Neural Networks. From their experimental results, they concluded that the determining diagnostic system could increase the amount of decision making during the prognosis process of coronary disease.
Sentil et al. [3] have proposed a hybrid random forest with a linear model(HRFLM) for providing enhanced work level with high efficiency through the prediction method for heart disease. The performance of Decision Tree, Naive Bayes, Support Vector Machine, Random Forest, and HRFLM are compared. They proposed HRFLM method gives better accuracy.

Binhua et al. [4] have extensively presented multiple tasks deep and wide neural network (MT-DWNN) applied to the dataset collected from the Chinese PLA General Hospital data. The researchers show that the MT-DWNN achieves better prediction performance on renal dysfunction in Heart Failure patients than conventional models.

Prakash et al. [5] did a study on coronary disease prediction and proposed Optimality Criterion Features selection (OCFS). OCFS is used for the extrapolation and proficiency diagnosis of coronary disease. The researchers used a rough set feature selection on information entropy (RFS-IE). In this study, they compare the OCFS with RFS-IE in terms of computational time, prediction quality, and error rate. The researchers claimed the OCFS technique takes less execution time when compared to other methods.

BACKGROUND RELATED WORK:

The researchers [1-5] using data mining and machining learning techniques [6-10]. These two approaches are fundamental to medical systems. Various supervised learning approaches like Decision tree, k-nearest neighbor, Naive Bayes, Random Forest, and Neural Networks can be used for the prediction of heart disease. Machine learning algorithms can be put profoundly in all fields of medication, from medication disclosure to clinical practices. Machine learning gained popularity in Health care applications after the digitization of medical records in the past few decades. Among different machine learning algorithms heuristic methods, handcrafted feature

extraction techniques supervised learning techniques, namely convolution neural networks widely used in medical analysis

Machine learning: Natural language processing is a part of Artificial Intelligence (AI). AI profoundly associates produce methods that can get from involvement. The approach that machine learning methods work is that they identify unknown patterns in data and construct methods. Then, they can produce accurate predictions for new data files that are fully advanced for the methods. This way, the machine come more knowledgeable through learning, so it can analyze patterns that are very hard or inaccessible for peoples to find by themselves. ML algorithms and techniques can operate massive datasets and form decisions and predictions. The simplified machine learning is depicted in Fig1. In this fig, the dataset, which consists of patient-related data, is pre-processed first. The pre-processing part is essential as it removes the data and provides it to be recycled by the ML methods. The method consists of one or more algorithms working together in a hybrid approach. The result of the technique is an output classifier. The classifier receives input data and produces whether the new patient is healthy or unhealthy based on the data.

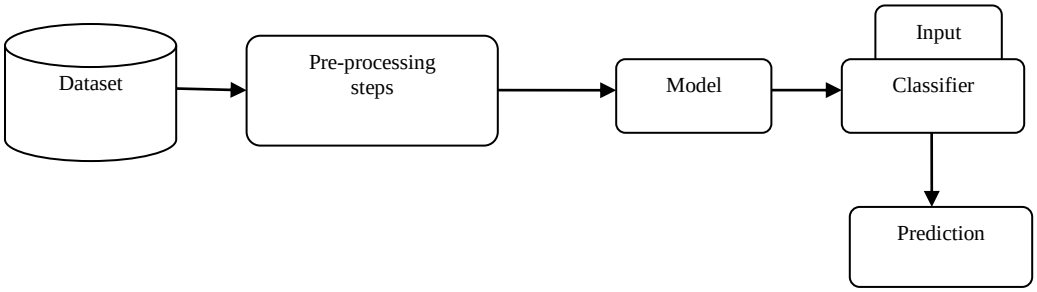

Figure 1. Simplified Machine Learning Diagram

III. PROPOSED METHOD

1) **K-nearest neighbor (KNN):** KNN is a non-periodic technique intended for classification and regression. KNN is also called a lazy algorithm. KNN technique is used for feature similarity to predict the values of any new data points. Euclidean distance is a distance measure that is widely used in KNN.

2) **Decision Tree:** Decision tree is a classification model, and it must be useful for the two classifications and regression, though it most popularly used for classification.

- **Decision nodes:** Decision tree we test something the test may have more than one result based on the value of the test either follow the branches. This test is usually done on attributes
- **Leaf nodes:** This indicates the classification or value of the example
- **Entropy:** It is estimated to dividing the input into groups consists of equal data containing identical values, and then the parent node is determined placed on the maximum information gain.
- **Information Gain:** After dividing the dataset based on the entropy change, information gain is estimated. The attribute which produces maximum if it is selected as a parent node.
- Entropy > 1, requires splitting
- Entropy = 0, indicates a leaf node

3) **Random Forest:** It is a supervised machine learning technique that constructs multiple decision trees. The random forest includes flexibility and converts high variance to low variance, and it grows decision trees from multiple bootstrapped samples. When training, each tree study from an element of the data points. The samples are substitute known as bootstrapping, which means in a single tree some samples are used various times. Each

decision is found by using a random subset of the training data. First, we select the unplanned elements from an inclined dataset, and the method builds up a decision tree for the whole sample. Formerly it will make the prediction outcome from the decision tree, and next voting will be implemented for every predicted outcome, Finally select the utmost chosen prediction outcome as the conclusion prognosis outcome

4) **Convolution Neural Networks:** In recent years, CNN enforced different classification work in the part of machine learning, such as text classification, and have been shown to have excellent classification ability. CNN is a kind of neural network, its weights distribution network structure generate it more identical to the biological neural network, decreases the intricacy of the network method, and the number of weights.
CNN has three types of layers they are shown below :
   i. Convolution layer
   ii. Pooling layer
   iii. Soft-max layer

In the convolution layer, input data is convolved with various kernels. CNN consistently uses spatial information and produce different component maps. The pooling layer diminishes the size of the feature map by spatial invariance average or highest activity. Both the convolution and pooling layer compose the feature extraction module. Convolution neural network is exploited as feature extraction. In soft-max layer, the soft-max activation function is used to classify the input feature map into class value

Implementation: By applying convolution operation to the data, CNN provides a visualization of an individual neuron. The pooling layer is given the output of this layer. In the next layer, in the pooling layer, the combination of neuron outputs is reduced to one single neuron. Hidden and output layers were initialized in completely connected layer data, and a sequential model was developed. There were 13 neurons in the input layer. Input parameters along with weights were provided to the neurons at the input layers, and these layers were trained by changing the Epoch value until the residual error was minimized. The Presence or Absence of Heart Failure Risk has been brought back to the output layer.

5) **Support Vector Machine:** This method has a useful order precision. It is defined as a low-dimensional vector space that comprises a measurement for each attribute of an object. An SVM[11] plays immediately and can be utilized as a prediction model. In other words, SVM is a model that uses machine learning theory to maximize predictive exactness while manually avoiding over-fit to the data SVM.

6) **Naive Bayes:** It is a supervised learning method utilized for classification tasks. Hence, it is also known as the Naive Bayes classifier. It uses features to predict the target variable. The key difference is that Naive Bayes considers that features are not dependent on one of each other, and there is no correlation among features. Naive Baye's classifier estimates the probability of a class given a set of feature values. The assumption that all features are independent reach naive Baye's model most quickly related to difficult methods. In a few cases, speed is taken over maximum exactness.
Scikit-learn Implementation: I will use one of the available datasets (heart disease) on sci-kit-learn. It is utilized to split the dataset into train and tests with the goal that we would first be able to prepare the model and afterward calculating the performance. Then we load the dataset and separated training and testing are in the ratio 70:30. Next we make a Gaussian Naive Bayes classifier object and fit train information to the classifier.

7) **Logistic Regression:** It is a machine learning technique, classifying records of a dataset based on the values of input. Logistic regression is basically a supervised classification algorithm.

8) **Gradient boosting Classifier:** Boosting is a technique of changing weak learners to strong learners. The gradient boosting algorithm can be mostly related by first presenting the AdaBoost algorithm. The AdaBoost starts with a decision tree in which each information is related to an equivalent weight. After classifying the primary node, we increase the weights of that information that is tough to classify and reduced the load for

those that are simple to classify. The secondary node is therefore grown on this weighted data. Gradient boosting trains several methods in a gradual, addictive, and continuous manner.

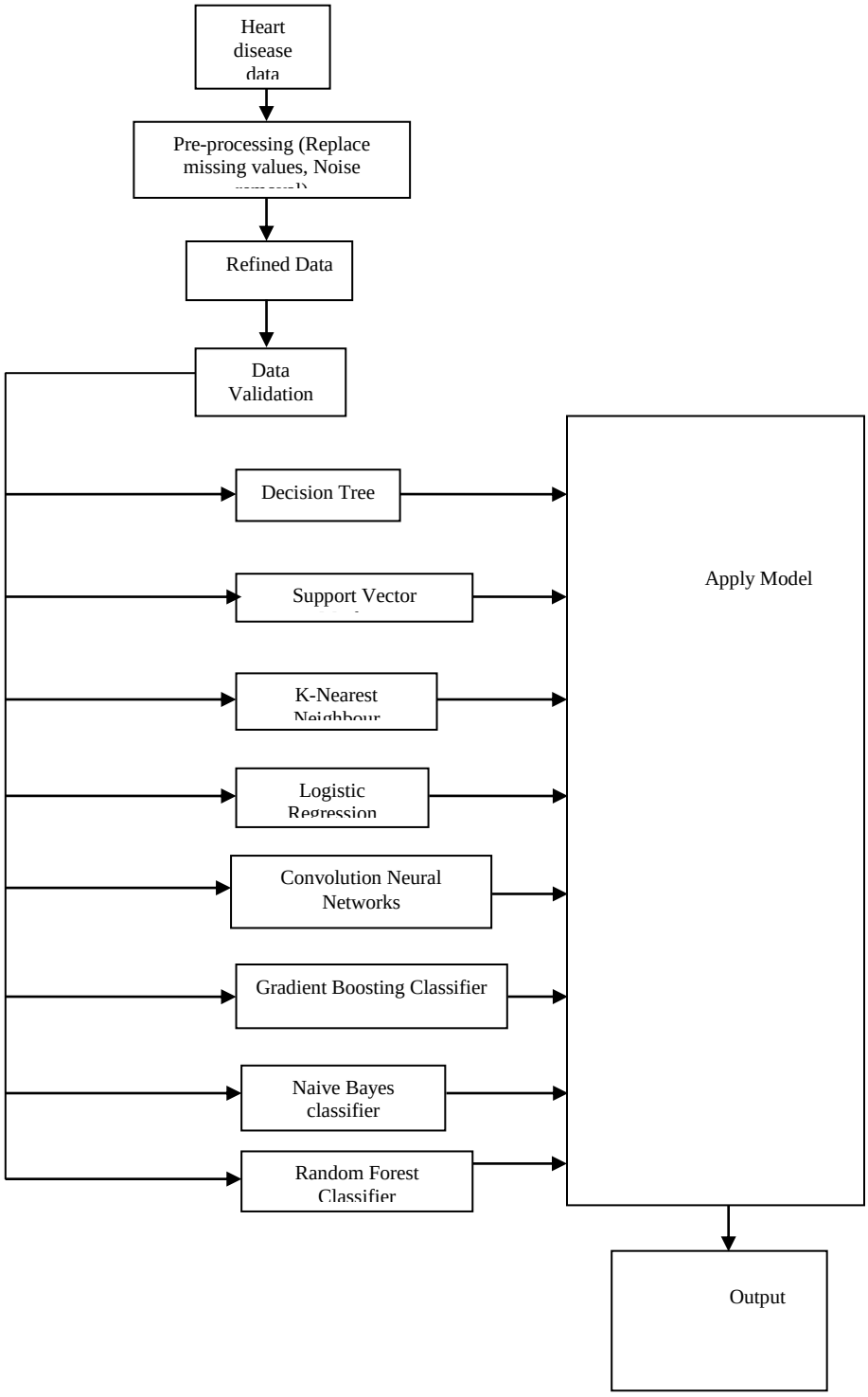

Figure 2. Proposed Architecture

III. DATASETS & EXPERIMENTAL RESULTS

Dataset:

In the study, we have used four datasets collected from the UCI repository (Cleveland, Hungarian, VA long beach, and Switzerland) is used to identify the Heart Failure problem in patients. The Cleveland data set consists of 303 tuples. The Hungarian data set consisting of 294 tuples. The VAlong beach data set consists of 200 tuples. The Switzerland data set consists of 123 tuples. 70% of tuples are used for training, and the remaining 30% of tuples are used for validation.

Experimental Observation:

In this study, online heart disease datasets such as Cleveland, Switzerland, Hungarian, VA long beach are available on UCI Machine learning Repository used for heart failure risk in patients. First, we load the data and applying data pre-processing steps to the dataset to make data more suitable. Fill in missing values, identify outliers and remove them, and resolve inconsistencies.

Experiments are conducted on the Clevelan data set. Using KNN Classifier produces 87% accuracy. SVM classifier produces 84% accuracy. Decision Tree classifier produces 86% accuracy. Random Forest Classifier produces 86% accuracy. The Convolution Neural Network classifier produces 88% accuracy. Logistic regression produces 88% accuracy. Gradient Boost classifier produces 92% accuracy. Naive Bayes Classifier produces 86% accuracy. The results are given in Table 1.

Experiments are conducted on the Switzerland data set. Using KNN Classifier produces 83% accuracy. SVM classifier produces 81% accuracy. Decision Tree classifier produces 83% accuracy. Random Forest Classifier produces 83% accuracy. The Convolution Neural Network classifier produces 87% accuracy. Logistic regression produces 86% accuracy. Gradient Boost classifier produces 92% accuracy. Naive Bayes Classifier produces 88% accuracy. The results are given in Table 1.

Experiments are conducted on the Hungarian data set. Using KNN Classifier produces 76% accuracy. SVM classifier produces 75% accuracy. Decision Tree classifier produces 77% accuracy. Random Forest Classifier produces 78% accuracy. The Convolution Neural Network classifier produces 88% accuracy. Logistic regression produces 93% accuracy. Gradient Boost classifier produces 91% accuracy. Naive Bayes Classifier produces 86% accuracy. The results are given in Table 1.

Experiments are conducted on the VALong beach dataset. Using KNN Classifier produces 76% accuracy. SVM classifier produces 75% accuracy. Decision Tree classifier produces 73% accuracy. Random Forest Classifier produces 76% accuracy. The Convolution Neural Network classifier produces 77% accuracy. Logistic regression produces 80% accuracy. Gradient Boost classifier produces 91% accuracy. Naive Bayes Classifier produces 86% accuracy. The results are given in Table 1.

Table - 1 Experimental results on Clevelan dataset, Switzerland dataset, Hungarian dataset and VA Long beach dataset

Techniques	Clevelan dataset	Switzerland dataset	Hungarian dataset	VA Long beach dataset
KNN	87%	83%	76%	76%
Support Vector Machine	84%	81%	75%	75%
Decision Tree	86%	83%	77%	73%
Random Forest Classifier	86%	83%	78%	76%
CNN	88%	87%	88%	77%
Logistic Regression	88%	86%	93%	80%
Gradient boosting classifier	92%	92%	91%	91%
Naive Bayes classifier	86%	88%	86%	86%

A comparative study has been performed among various classifiers on different datasets. The artifacts are presented in figure 3.

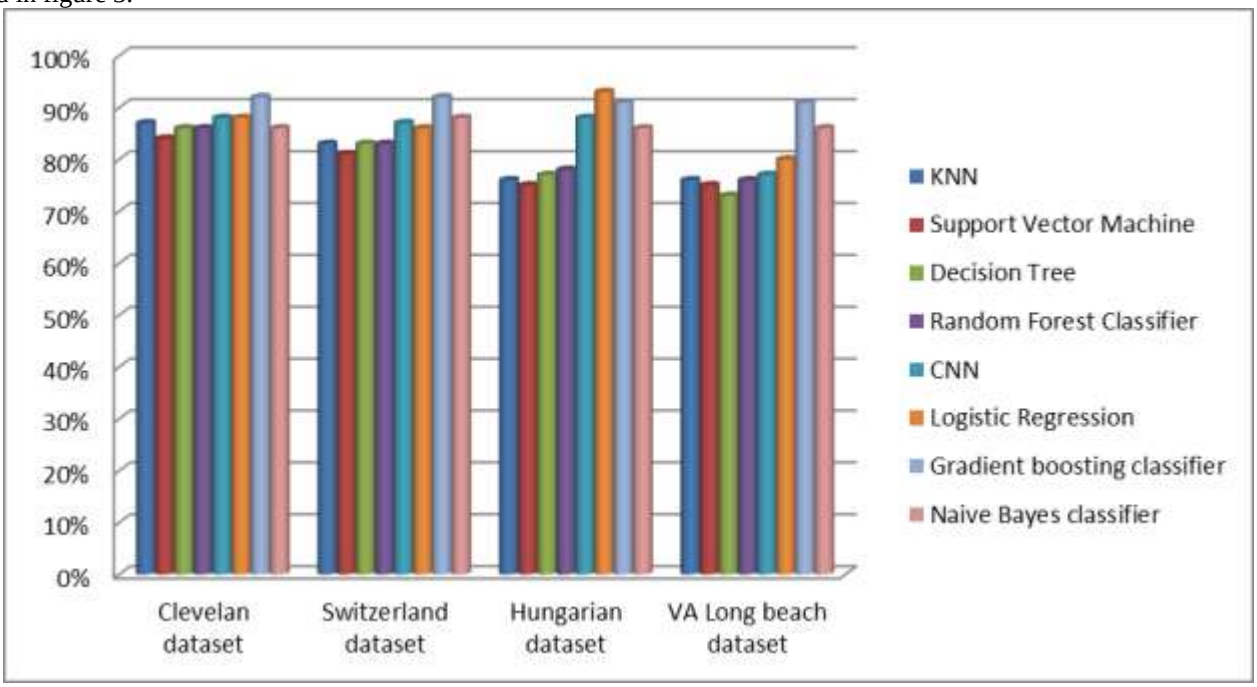

Figure 3. Comparative study of classifiers on different datasets

IV. CONCLUSION

The cardio disease is accompanied by numerous serious out-comes in particular, consistent hospitalization, increased transience, high monetary loss, and poor quality of life. We provide a comparative study of data mining and machine learning techniques, such as KNN, Support Vector Machine, Decision Tree, Random Forest, Convolution Neural Network, LogisticREgression, Gradient boosting, Naïve Bayes classifiers. Experimental results show that the Gradient boosting classifier is giving the best outcome as compared to other methods.

REFERENCES

[1] Tiwaskar, Rutuja Gosavi, Riddhima Dubey, Shaila Jadhav, and Komal Iyer "Comparison of Prediction Models for Heart Failure Risk : A Clinical Perspective", Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), 978-1-5386-5257-2/18/$31.00 ©2018.IEEE.

[2] Liaqat ali , atiqur rahman, Aurangzeb khan, mingyi zhou,ashir javeed, and javed ali khan "An Automated Diagnostic System for Heart Disease Prediction Based on _2 Statistical Model and Optimally Configured Deep Neural Network", Access IEEE, vol. 7, pp. 180235-180243, 2019

[3] Senthilkumar mohan, chandrasegar thirumalai,and gautam srivastava," Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques, Access IEEE, pp. 81542-81554, 2019. .Digital Object Identifier 10.1109/ACCESS.2019.2923707,volume7 2019,81543

[4] Binhua wang , yongyi bai, zhenjie yao, jiangong li, wei dong,yanhui tu, wanguo xue , yaping tian, yifei wang, and kunlun he ," A Multi-Task Neural Network Architecture for Renal Dysfunction Prediction in Heart FailurePatients With Electronic Health Records", in *IEEE Access*, vol. 7, pp. 178392-178400, 2019, doi: 10.1109/ACCESS.2019.2956859.

[5] Prakash, S., Sangeetha, K. & Ramkumar, N. An optimal criterion feature selection method for prediction and effective analysis of heart disease. Cluster Comput 22, 11957–11963 (2019). https://doi.org/10.1007/s10586-017-1530-z

[6] Hemantha Kumar Kalluri, MVNK Prasad, A Agarwal, Palmprint identification based on wide principal lines, Proceedings of the International Conference on Advances in Computing

[7] Sajja Tulasi Krishna, H K Kalluri, Deep Learning and Transfer Learning Approaches for Image Classification, International Journal of Recent Technology and Engineering (IJRTE)

[8] Tulasi Krishna Sajja, Retz Mahima Devarapalli, H K Kalluri, Lung Cancer Detection Based on CT Scan Images by Using Deep Transfer Learning, Traitement du Signal 36 (4), 339-344

[9] VR Sajja, HK Kalluri, Brain Tumor Segmentation Using Fuzzy C-Means and Tumor Grade Classification Using SVM, Smart Technologies in Data Science and Communication, 197-204

[10] ST Krishna, HK Kalluri, Lung Image Classification to Identify Abnormal Cells Using Radial Basis Kernel Function of SVM, Smart Technologies in Data Science and Communication, 279-285

[11] V Ramakrishna Sajja, Gnaneswara Rao Nitta, "Experimental Approache for Detection of Brain Tumor Grade Using Svm Classification", International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-7, Issue-5S4, February 2019, Retrieval Number: E10160275S419/19©BEIESP, PP.79-85