Research Article

International Journal of Advanced Computer Research, Vol 10(49)
ISSN (Print): 2249-7277 ISSN (Online): 2277-7970
http://dx.doi.org/10.19101/IJACR.2020.1048064

Predictive and perspective analysis of cancer image data set using machine learning algorithms

Divya Chauhan[1*] and Kishori Lal Bansal[2]
Assistant Professor of Computer Applications in Government College Rampur, Himachal Pradesh, India[1]
Professor, Department of Computer Science, Himachal Pradesh University, Shimla India[2]

Received: 14-May-2020; Revised: 22-July-2020; Accepted: 24-July-2020

Abstract
Classification and prediction of the images are fairly easy task for humans, but it takes more effort for a machine to do the same. Machine learning helps to attain this goal. It automates the task of classifying a large collection of images into different classes by labelling the incoming data and recognizes patterns in it, which is subsequently translated into valuable insights. The aim of this paper is to classify the image data set of five cancer types, namely Osteosarcoma, Prostate Cancer, Brain Cancer, Breast Cancer and Acute Myeloid Leukaemia. Furthermore, the prediction of Osteosarcoma case for one of the four classes of tumor namely Non tumor, Non-Viable tumor, viable tumor, Viable: Non-Viable tumor has to be done. The quantitative analysis is done using various machine learning libraries of python. The three classification algorithms used for image analysis are random forest, SVM, and logistic regression. The metrics used for performing perspective analysis are precision, recall and F1 Score. The results show that the random forest algorithm has performed best amongst the three classification algorithms when given with less complicated scenario, with prediction accuracy, precision, recall and f1 score of 100%. But the performance of every classification algorithm degrades when provided with the cases of Osteosarcoma which has got more complicated scatter graph. However, the logistic regression retains its performance by predicting tumor cases with 99% accuracy.

Keywords
Data mining, Big data, Hadoop, Mahout, Clustering, Health care.

1.Introduction
Undoubtedly, machine learning is becoming a daily reality and need of the hour. It is an application of artificial intelligence that provides a system the ability to automatically learn as well as improve from the past experience. In fact, it deals with the learning process in which machine tends to learn on its own without being explicitly programmed [1]. With time, machine learning has evolved by gaining new momentum as a consequence of learning from big data. Evidently it has brushed up the ability to automatically apply complex mathematical statistics on big data with greater efficiency and speed. Nowadays, the techniques of machine learning are ushering in mental and cognitive ability and can change the world if used deftly and calculative to harness its power.

Besides using the ability of machines to store and access more data than a person and adding machine learning on top of it to identify trends leads to arrive at a solution to previously untenable problems.

One of the biggest applications of big data and machine learning is in the field of medical domain. Consequently, a health care organization that uses the techniques of machine learning and big data to treat patients see fewer mishaps or gets enough time to deal with them in advance. It is also helping medical organisations to tackle some of the most intractable problems by allowing the researcher to better understand the disease and predict the outburst of disease through the use of predictive models [2].

There are many different kinds of machine learning algorithms to discover certain patterns in big data that leads to actionable insights. At the broader level, these machine learning algorithms can be divided into two groups based on the way they learn from the data to make predictions. These two groups are

*Author for correspondence

supervised algorithm and unsupervised algorithm. The one used in this paper is supervised machine learning algorithms.

1.1Supervised machine learning

The learning model is provided with input variable as well as correct target variable. In addition, the model learns the mapping function and trained to the extent where it gains the ability to predict the class of unknown variable. Supervised algorithms are further classified into regression and classification. Both classes of supervised learning strive to construct an efficient learning model that can predict the class of unknown variables accurately [3]. It is important to understand which algorithm is to be used depending upon the application domain and the type of data provided to the learning model.

Regression algorithm are used when the variables has real or continuous values. Regression can be linear or nonlinear, in addition, can be simple with one feature or multiple with more than two features in the output variable. The regression model maps the input space into a real value domain.

Classification algorithms are used when the output variable is categorical in nature. Few use cases of classification are document classification, handwriting recognition, speech recognition, biometric regression and many more [4].

The three classification algorithms used and discussed in this paper are as follows:

1.1.1Logistic regression

It is a supervised learning algorithm in which target variable can only take discrete values from a given set of features. It is a regression model which predicts the probability that given data belongs to which particular category of output/ target variable [5]. It models the data using sigmoid function g(z)=1/(1-e^(-z)). Logistic regression can be binomial (yes and no), multinomial (10, 20, 30, …) or ordinal (Sunday, Monday,..). One assumption of logistic regression is that the data should not have the problem of multi collinearity that is the independent variables should be independent of each other. Accordingly, this algorithm works well with larger data set in comparison to smaller one.

1.1.2Random forest

This algorithm is used for both regression and classification problems. It is an ensemble learning method which has proven to be better than single learning algorithm. It creates decision tree on the given data set and gets the prediction for each one of them and eventually selects the best solution by the means of voting process. Subsequently the algorithm reduces the problem of over fitting by averaging the result. The working of random forest algorithm is quite simple indeed. It starts with the selection of random samples from the given training set. Next it constructs the decision trees for every sample and then it gets the prediction result for every decision tree. The best prediction is selected by voting process and outputted as final prediction result. The algorithm works well with large data set as it has less variance. Besides it is quite flexible and possesses high accuracy even after providing the data without scaling and missing values, but is time consuming than other learning algorithms [6].

1.1.3Support vector machine

It is able to handle multiple, continuous and categorical variables. SVM creates a model which is basically the representation of different classes in a hyperplane with multi-dimensional space. The hyperplane is generated in an iterative manner by SVM so that the error can be minimized. The goal of SVM model is to divide the given data set into classes in such a way that maximum marginal hyperplane can be found. The algorithm needs to implement following two steps: firstly, segregation of classes is required which needs the formation of such hyper plane iteratively. Secondly the best hyperplane will be selected which separates the classes correctly. Evidently the Support vector machine algorithm works efficiently with high dimensional space with great accuracy [7]. It uses subsets of training data passes to it and hence requires less memory space for processing. As the time required to train the SVM model is high so it is not suitable for large data sets and it also shows poor performance with overlapping classes.

1.2Problem statement and objectives

It is a fact that human can visualize the images easily, in contrary it may seem a problem for the machines. However, a rigorous research and development for machine learning techniques has been evolved to improve imaging techniques, principles, extension of computer power and space, but there seems a need to use them efficiently in the medical domain for medical diagnosis, treatment planning, etc. Though, there is also a certain difference in the range/ tolerance for the precision of accuracy of each problem, where it can be accepted clinically. This paper focuses on the problem of using machine learning techniques on cancer/tumor prediction.

Objectives:

The main objectives of the work are as follows:

1. To classify the image dataset into five different cancer class using three classification algorithms.
2. To predict the class of tumor in one of the cases of cancer.
3. To perform perspective analysis on the features extracted from the cancer image dataset.

2.Literature review

Nadiammai and Hemalatha [8] applied data mining algorithms to intrusion detection databases. Various rule-based classifier like Ridor, Conjuctive rule, One R, Part, Zero R, DTNB, NNge, Decision table, RBF, multilayer perception and SMO are used.10 fold cross validation has been used for verification. Accuracy, sensitivity and specificity are calculated to measure the performance in which SMO classification algorithm performed best among all. For future scope hybrid detection system can be developed based on data mining algorithms. Liu et al. [9] presented a comprehensive analysis and interpretation of Artificial Intelligence, big data and data mining. Various research opportunities are also highlighted. Khatavkar et al. [10] performed multi perspective analysis on BBC news using machine learning algorithms named decision tree, random forest, AdaBoost and SVM. Various standard metrics for analysis are Kappa, f1 score, recall, accuracy and precision. SVM with linear svc gave classification rate of 96% and positive rate of 75%. Celli et al. [11] classified large DNA Methylation data set to identify cancer drivers. They proposed BIGBIOCL algorithm which could apply machine learning algorithms on hundreds of thousands of features in few hours. Performance of BIGBIOCL is compared with state of art classifiers. Khalifa et al. [12] proposed a framework called LADEL and implemented it on Apache Drill to distribute the training execution of Weka's classification algorithms. Results showed that LADEL distributed classifiers have similar and sometimes even better accuracy to the single-node classifiers and they have a significantly faster training and scoring times. Genevès et al. [13] predicted patients at risk by analysing drug prescription data using binary classification model. The data comprised of millions of patients and hundreds of hospitals. Classifiers included were decision trees, random forest, SVM. Sun et al. [14] implemented Lossless Pruned Naive Bayes (LPNB) classification algorithm to classify thousands of classes. Results showed that for real-world data set with 7205 classes, LPNB can classify text up to eleven times faster than standard Naive Bayes. McGinnis et al. [15] used wearable sensors and machine learning to diagnose anxiety and depression in young children. The study demonstrated that 20 seconds of wearable sensor data extracted from a fear induction task, when combined with machine learning, can be used to diagnose internalizing disorders in young children with a high level of accuracy and at a fraction of the cost and time of existing assessment techniques. Dumitrescu et al. [16] proposed Penalized Logit Tree Regression (PLTR) to improve the framework of logistic regression by using information from decision trees. Rules were extracted from various short-depth decision trees built with different sets of predictive variables and are used as predictors in a penalized or regularized logistic regression. The proposed algorithm is implemented on real world data for credit scoring and the performance over performed the traditional random forest and traditional logistic regression. Xin and Wang [17] proposed a training criterion of depth neural network for maximum interval with minimum classification error by combining the cross entropy and M3CE. The testing of proposed method on two deep learning standard databases showed positive results by enhancing the cross-entropy. Gupta [18] has explored existing and future opportunities in the field of image processing and computer vision. Many data repositories, application of each band of the electromagnetic spectrum and image examples are illustrated. Bianco et al. [19] presented a novel CNN- based method combined with spline-based color curves to estimate a global color transformation for raw image enhancement in order to improve the perceived quality of images. It is also used to model multiple experts at the same time to not incur any accuracy loss or computational resources. Liu et al. [20] proposed a loss function named refocused image error to optimize the image quality of synthesized light field in refocused image domain, subsequently, the performance is tested against previous approaches on both real and software rendered light field datasets. Liu et al. [21] introduced a deep learning method for image quality assessment for paediatric and weighted MR images, which is initially performed slice-wise and then volume-wise using random forest, subsequently testing it exhibit great generalization with near-perfect accuracy. Yasarla et al. [22] proposed a novel multi-stream architecture called Uncertainity guided Multi-stream Semantic Network (UMSN) and training methodology to exploit semantic labels for facial image deblurring using predicted confidence measure during training in a end to end fashion. Significant improvements are noticed when evaluated for three different face datasets.

3.Results and analysis

3.1Data set used

The image data set for classification is retrieved from TCIA for cancer imaging archives. This site is funded by the National Cancer Institute's (NCI) cancer imaging program. TCIA is a service which hosts a large archive of many medical image of cancer accessible for the researcher to download. The data is categorized as the collection of different cancer cases, image modality or research focus. The majority of data consists of CT, MRI and nuclear medicine images. The primary file format used by TCIA images is DICOM. DICOM stands for Digital Imaging and Communications in Medicine. It is an international standard to store, transmit and process medical imaging information. *Figure 1* below shows the samples of five types of cancer cases which are considered for image classification and subsequently calculating the prediction accuracy of learning models.

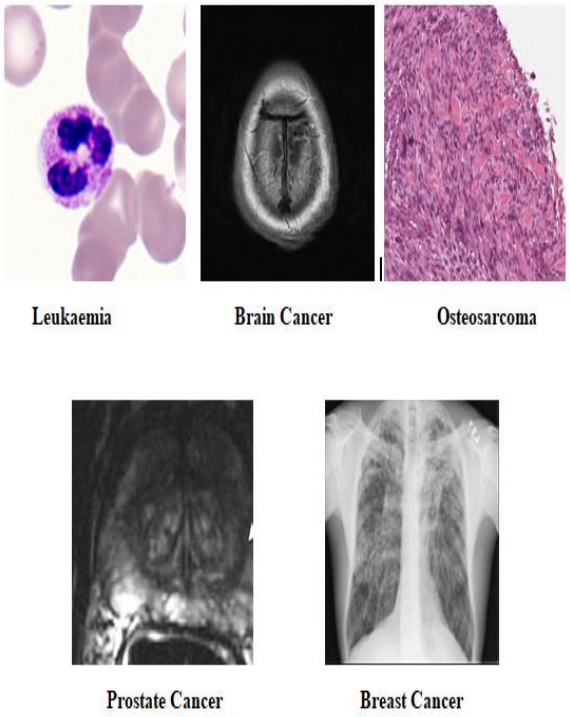

Figure 1 Five cancer types used for analysis

3.2Process block diagram

The main steps carried out in this work are shown below in *Figure 2* in the form of block diagram and are subsequently elaborated in coming sections.

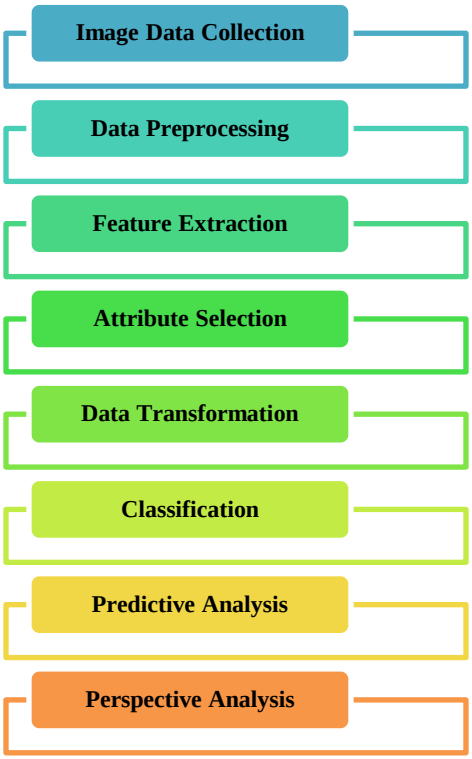

Figure 2 Block process diagram

3.3Preprocessing of data set

Generally, machine learning is completely dependent on the data set being used to train the model. There is always a requirement of the right data that is properly scaled, formatted and contains meaningful features. Therefore, data preprocessing serves a very important role in machine learning. It converts the data in the form that is required to be fed into the machine learning algorithms to get the model work as per expectation. An image is nothing but a two-dimensional array of numbers ranging between 0 and 255. It is defined by two coordinates x and y for horizontal and vertical position respectively. The value (x, y) at any point gives the pixel value at that particular point of an image. The images downloaded for the analysis are in different file formats like some of the images are downloaded in the DICOM file format. Subsequently, NBIA data retriever is needed to access this file format which is open source software. With the help of NBIA data retriever, the file gets downloaded in .dcm file format which is still incompatible to be fed into machine learning algorithms. Likewise, PearlMountain Image Converter comes into play for converting .dcm files into jpeg or other compatible file formats. The file format used in this work is .jpg, .png and .tiff to get the variety into the data set.

One of the tasks in preprocessing is feature selection of attributes provided to the machine learning algorithm where the performance of the machine learning algorithm is directly proportional to the feature selected in the training set. By the same token, the performance of algorithm will be negative if the data with irrelevant features are fed to the learning process and quite positive when the data features fed are relevant. The selection of relevant features is also known as attribute selection. Few advantages of attribute selection are reduction in processing time, reduction in overfitting and increase in accuracy of the model. Python is used for digital image processing for allowing much wider range of algorithms to be applied to the input data; accordingly, it improves the image data or features by suppressing unwanted noise and enhances some important image features so as the machine learning algorithms can build a better model.

Finally, the data is transformed into a similar type so that different data can be processed altogether. As we are aware that Pandas library has many techniques that make the process of extraction, filtration and transformation of data efficient and intuitive. Pandas data frame is a two-dimensional size mutable, potentially heterogeneous tabular data structure with labelled axis. It allows the arithmetic function to be performed on both row-wise and column-wise and can be thought of as a dictionary like container for series objects. There are various types of transformation used in this paper, one of which is scaling. Surely scaling plays an important role when the data is fed to certain machine learning algorithms. Nonetheless, the data set collected from different sources comprises of attributes with varying scale but that must be rescaled to make sure that attributes are at same scale. Generally, data is normalized into the range of 0 and 1. Scikit learn library of python is used for this purpose with the following syntax:
preprocessing.MinMaxScaler(feature_range-(0,1)).
fit_transform(array)

So the major steps for preprocessing the image data sets are as follows:

- Convert the DICOM files into compatible file formats like .jpg, .png and .tiff.
- Read the images into arrays.
- Getting the desired features of all images by resizing, feature selection, transformation and scaling of image data.

3.4Classification of data

The data set comprises of five types of cancer cases images which are Osteosarcoma, Prostate Cancer, Brain Cancer, Breast Cancer and Acute Myeloid Leukaemia. The number of occurrences of individual cancer cases in the data set is shown in *Figure 3* below in the form of pie chart.

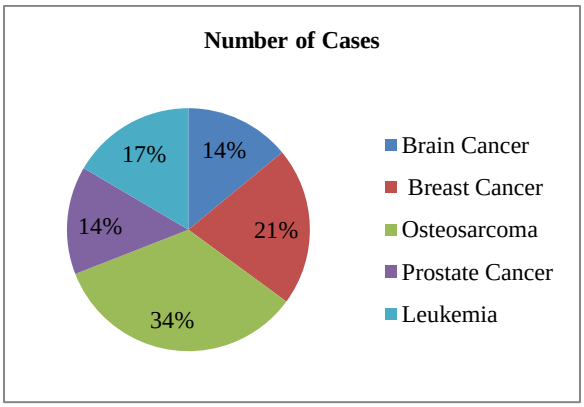

Figure 3 Five cases type in cancer data set

Since Osteosarcoma is having the highest occurrences in the data set, it is chosen to further extract features and categories each image into one of the four categories namely Non Tumor, Non-viable Tumor, Viable tumor, Viable: Non-viable tumor.

- Non tumour is a class that is not malignant; they grow larger but do not spread to other parts of the body.
- Non-Viable tumor is dead and unable to grow.
- Viable tumor is defined as the presence of epithelial cells within the lymph node. It can grow and divide or develop.
- Viable: non-viable tumor is class where it is not clear to which category the tumor actually belongs to.

Figure 4 below shows the number of cases of each tumor category of Osteosarcoma cancer present in the data set.

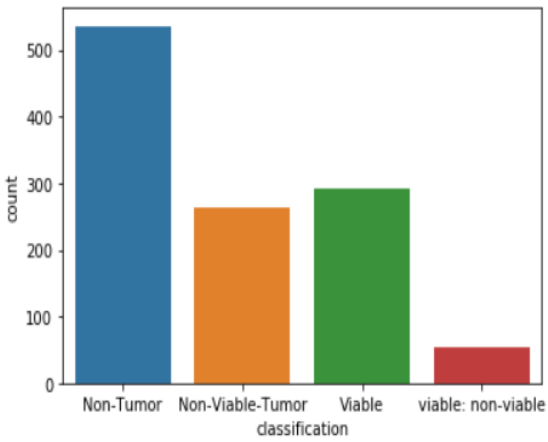

Figure 4 Four tumor classes of Osteosarcoma

3.5Splitting data set

There are two major problems while generating the learning model from the data using any machine learning model. They are overfitting and underfitting. Overfitting refers to the model which has been trained too much and fits closely to the training data set. This model is not generalized and cannot classify other unknown data. This usually happens either the model is too complex or the number of features is more compared to the number of observations in the data set. Consequently, the model shows high accuracy with the training data set but will not be able to classify the test set with that same accuracy. One of the reasons is that model learns on noise instead of the actual relationship between variables present in the data.

Underfitting refers to the inability of a model to fit in the training set and therefore missing the trends in the data causing the model not to generalize to new unlabelled data. One of the reasons is not having enough predictions in the data set and the other reason can be trying to fit nonlinear data in the linear model.

For the sake of avoiding both of the above-mentioned problems in data modelling, a middle ground needs to be chosen between overfitting and underfitting the model. Certainly, learning the model on a data set and testing it on that same data set is a methodological mistake. As a consequence of above-mentioned scenario, the trained model tends to repeat the labels of the samples that have already been seen and presents a perfect score in learning but is not able to predict anything useful on unknown/ unlabelled test data set.

Therefore, for each algorithm, the data set is divided into two sets. One set for training the learning model called training set and another for validation and testing the model which is called test set. Stratified random split is used for partitioning the two data sets so that all same types of classes do not fall into the single set. *Figure 5* below shows the code to use stratified random split using python.

```
split=StratifiedShuffleSplit(n_splits=1,test_size=0.2,random_state=50)
for train_index,test_index in split.split(X,y):
    X_train_set=X.loc[train_index]
    X_test_set=X.loc[test_index]
    y_train_set=y.loc[train_index]
    y_test_set=y.loc[test_index]
```

Figure 5 Stratified shuffle split

3.6Predictive analysis

There are three classification algorithms that classify five types of cancer cases and identify the class of tumor amongst the four classes. These three classification algorithms are:

- Random forest
- Support Vector machine
- Logistic Regression

Python uses the Scikit learn library to import certain functions for machine learning algorithms. For random forest, it imports RandomForestClassifier and uses the method as follows:
model_name=RandomForestClassifier(attributes)
model_name.fit(X_train_set,y_train_set)
For Support vector machine it uses:
model_name= SVC(attributes)
model_name.fit(X_train_set,y_train_set)
and for using logistic regression for building learning model it imports LogisticRegression and uses method as follows:
model_name=LogisticRegression(attributes)
model_name.fit(X_train_set,y_train_set)
where *attributes* is replaced by certain properties which needs to be changed for the particular machine learning algorithm.
Accordingly, the learning model is trained using this training set using each of the algorithms separately. Each learning model is fed with the test data set to check the accuracy of built model. Likewise, the two tables constructed below show the accuracy of all three-classification algorithm to classify different cases of cancer and also shows the number of cases the learning model has misclassified.
Table 1 shows the number of misclassified cases and prediction accuracy.

Table 1 Accuracy and misclassified cancer cases

Accuracy of cancer cases		
Algorithm	**Accuracy**	**Misclassified**
Random Forest	1.00	0
Logistic Regression	0.98	2
Support Vector Machine	0.99	1

Clearly all the algorithms are performing quite well with good prediction accuracies. It is evident that the random forest has learned quite well and is best in predicting the accuracy of cancer class with 100% accuracy. Secondly, SVM has one misclassified cases of cancer giving the accuracy of 99%. In the end, the logistic regression identifies two test cases of cancer incorrectly giving the prediction accuracy of 98%.
For visualization, the scatter plot of different cases of cancer is shown in *Figure 6* below

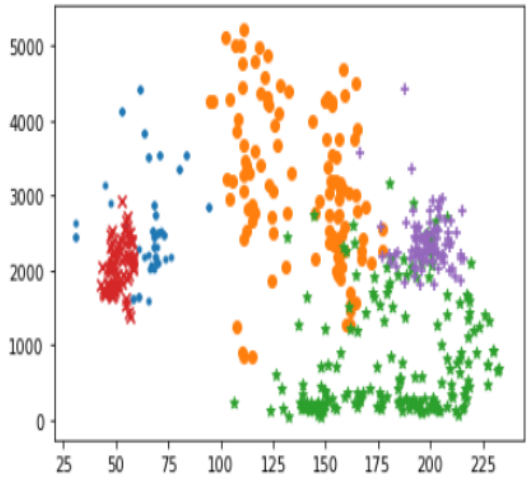

Figure 6 Scatter plot of cancer cases

Here red indicates Prostate Cancer, blue indicates Brain Cancer, orange color indicates Breast Cancer, purple plus sign indicates Leukaemia, and green color indicates Osteosarcoma.
In the same way, the three algorithms are applied to the training set to learn the tumor class categories in Osteosarcoma.

Table 2 shows the prediction accuracy of Osteosarcoma which has been divided into four classes of tumour. Here logistic regression is performing best with 99% accuracy and only two misclassified items whereas the random forest algorithm is able to predict tumor classes with an accuracy of 82% giving 63 wrong results. However, SVM is performing poorly with 158 misclassified items degrading its accuracy to only 54%.

Table 2 Accuracy and misclassified tumor classes

Accuracy of cancer class		
Algorithm	**Accuracy**	**Misclassified**
Random Forest	0.82	63
Logistic Regression	0.99	2
Support Vector Machine	0.54	158

Figure 7 below shows the scatter plot of tumor cases of Osteosarcoma where blue plus sign represents non tumor case, red color shows viable: on viable tumor cases, orange color shows non-viable cases and finally green dot represents viable tumor cases of Osteosarcoma.

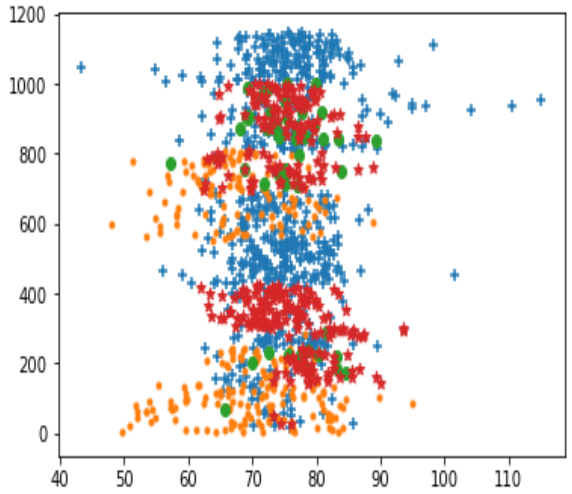

Figure 7 Scatter plot for four tumor classes

3.7Machine learning perspective analysis

Additionally, various evaluation metrics are considered for perspective analysis. There are several methods available to evaluate the learning model [23]. Some of them used are: precision, recall, F1 score, confusion matrix. *Table 3* below shows the evaluation parameter calculated for all three classification algorithms for predicting the five classes of cancer cases. It is quite clear that the random forest algorithm shows perfect precision, recall and F1 score with good support. Likewise, logistic regression has a perfect score for brain cancer, breast cancer and prostate cancer. Overall, the performance of classifier is good. Finally, for support vector machine the evaluation matrix is better if compared with the evaluation matrix of logistic regression.

Table 4 describes the evaluation matrices of cancer classes for three classifiers. It has been observed that logistic regression is performing best among all three

classifiers with an average of 99% of precision, recall and f1 score and support of 164. Random forest comes after that with 81% precision, 83 % recall and 82% f1 score, as the classifier lags in predicting viable tumor cases. On the other hand, SVM is performing poorly with an average precision of 40%, recall of 54% and f1 score of 45%. It is unable to classify viable and non-viable cases of tumor.

Table 3 Evaluation criteria of Cancer cases for three classifiers

Random Forest				
Cases	**Precision**	**Recall**	**F1-Score**	**Support**
Brain Cancer	1	1	1	16
Breast Cancer	1	1	1	24
Osteosarcoma	1	1	1	38
Prostate Cancer	1	1	1	16
Leukaemia	1	1	1	18
Weighted Average	1	1	1	112
Logistic Regression				
Brain Cancer	1	1	1	16
Breast Cancer	1	1	1	24
Osteosarcoma	0.97	0.97	0.97	38
Prostate Cancer	1	1	1	16
Leukaemia	0.94	0.94	0.94	18
Weighted Average	0.98	0.98	0.98	112
Support Vector Machine				
Brain Cancer	1	1	1	16
Breast Cancer	1	1	1	24
Osteosarcoma	0.97	1	0.99	38
Prostate Cancer	1	1	1	16
Leukaemia	1	0.94	0.97	18
Weighted Average	0.99	0.99	0.99	112

Table 4 Evaluation criteria of Tumor class for three classifiers

Random Forest				
Class	**Precision**	**Recall**	**F1-Score**	**Support**
Non Tumor	0.88	0.90	0.89	171
Non-viable Tumor	0.78	0.84	0.81	70
Viable	0.20	0.05	0.08	19
Viable: Non-Viable	0.83	0.87	0.85	83
Weighted Average	0.81	0.83	0.82	343
Logistic Regression				
Non Tumor	1	1	1	1
Non-viable Tumor	0.97	1	0.99	39
Viable	1	0.82	0.90	11
Viable: Non-Viable	0.97	1	0.98	29
Weighted Average	0.99	0.99	0.99	164
Support Vector Machine				
Non Tumor	0.55	0.83	0.66	166
Non-viable Tumor	0.00	0.00	0.00	81
Viable	0.00	0.00	0.00	9
Viable: Non-Viable	0.52	0.55	0.54	87
Weighted Average	0.40	0.54	0.45	343

4.Discussion

This section gives the overview of main comparisons extracted from the results section.

Comparison of scatter plots of cancer cases and tumor cases

The scatter plot constructed for five cases of cancer image dataset and four classes of tumor, gives the clear visualization of the dataset. Subsequently, it is observed that the scatter plot representing the different cancer class gives much distinguishable

classes as compared to the scatter plot representing the classes of tumor which is a chockablock of colorful symbols referring to different tumor classes.

Comparison of classification among different cancer cases

The three classification algorithms used for the classification of five different classes of cancer present in the image data set are random forest, logistic regression and SVM algorithm. However, all the three classification algorithms are giving good prediction accuracy but the random forest algorithm scores the best and the logistic regression algorithm scores the least in prediction accuracy.

Comparison of classification among different tumor cases

The same classification algorithms are used for the classification and prediction of four classes of tumor, providing the features extracted from the images. While comparing the three algorithms, logistic regression scores good enough but the prediction accuracy of random forest degrades to a lower score, and SVM scores below average.

Perspective analysis for cancer image dataset

The four evaluation criteria used for the perspective analysis of image dataset are precision, recall, F1 score and support. While dealing with a little less complicated cancer image data set, where the number of images are moderate and identical, three cases which have black and white images namely brain cancer, breast cancer and prostate cancer has scored perfect whereas the random forest and SVM gives moderate score for Acute Myeloid Leukaemia and Osteosarcoma. When dealing with a complicated cancer image data set, with large number of fairly identical images, the score of SVM is very poor, while non tumor, Non-viable tumor and Viable: Non-viable tumor have moderate score and viable tumor is hard to be identified by all three classification algorithms.

5.Conclusion and future work

The techniques of machine learning have been receiving a lot of attention in the world of analytics for many years. There are many different kinds of machine learning algorithms to discover patterns in big data that leads to actionable insights. Consequently, great demand of machine learning is sensed to increase the efficiency and insights of big data, especially, where the data is in unstructured formats and in a large amount. The domain of medical imaging helps providing important information on anatomy and organ function subsequently detecting disease states. Although the characteristics of medical data make its analysis a big challenge notwithstanding that machine learning techniques could make the task easier.

This paper uses three classification algorithms for analysis of five types of cancer cases namely Osteosarcoma, Prostate Cancer, Brain Cancer, Breast Cancer and Acute Myeloid Leukaemia. Furthermore, Osteosarcoma has been analysed to detect one of the four cases of tumors that are Non-Tumor, Non-viable Tumor, Viable, Viable: Non-viable. Moreover, to avoid the problems of overfitting and underfitting, separate training and test data set is provided to the learning model. Subsequently, for the perspective analysis three matrixes are constructed that are precision, recall, f1Score.

Ordinarily, every classification algorithm is performing sufficiently well when classifying as well as predicting 5 types of cancer cases. It has been observed that the best performance is given by random forest with prediction accuracy, precision, recall and f1 score of 100% each. Following that SVM has one misclassified cases of cancer giving the prediction accuracy of 99%. Finally, logistic regression identifies two test cases of cancer incorrectly giving the accuracy of 98%. One of the reasons is that the features extracted from the images of cancer cases are quite different from each other consequently easier to predict, which is quite clear from the scatter diagram itself. However, when a complicated application is fed to the same classifiers, the performance of classification algorithms degrades eventually. Consequently, the Support Vector Machine algorithm becomes the least suitable classifier with approximately 50% prediction accuracy to predict the cases of tumor. There is not a single case of viable and non-viable tumor which is predicted correctly. Logistic regression however performs well with 99% accuracy and only two misclassified items. Moreover, the Random forest classifier lags in predicting viable tumor cases of Osteosarcoma classes consequently winding the prediction accuracy to 82 % resulting overall 63 wrong results.

Overall, it can be concluded that the random forest classification algorithm is best suited when the problem is less complicated but logistic regression has performed best in this application domain with complex data set having very minute differences. For future scope, various other machine learning algorithms can be applied to observe their performance on the same set of features extracted from the cancer image data set.

Acknowledgment
None.

Conflicts of interest
The authors have no conflicts of interest to declare.

References
[1] https://searchbusinessanalytics.techtarget.com/ehandbook/Machine-learning-technology-techniques-add-new-analytics-smarts. Accessed 11 April 2020.
[2] Asim M, Khan Z. Mobile price class prediction using machine learning techniques. International Journal of Computer Applications. 2018; 179(29):6-11.
[3] https://towardsdatascience.com/a-brief-introduction-to-supervised-learning-54a3e3932590. Accessed 11 April 2020.
[4] Kesavaraj G, Sukumaran S. A study on classification techniques in data mining. In fourth international conference on computing, communications and networking technologies 2013 (pp. 1-7). IEEE.
[5] Korkmaz M, Güney S, Yiğiter ŞY. The importance of logistic regression implementations in the Turkish livestock sector and logistic regression implementations/fields. 2012; 16(2):25-36.
[6] Biau G. Analysis of a random forests model. The Journal of Machine Learning Research. 2012; 13(1):1063-95.
[7] Tong S, Koller D. Support vector machine active learning with applications to text classification. Journal of Machine Learning Research. 2001:45-66.
[8] Nadiammai GV, Hemalatha M. Perspective analysis of machine learning algorithms for detecting network intrusions. In third international conference on computing, communication and networking technologies 2012 (pp. 1-7). IEEE.
[9] Liu S, Wang X, Liu M, Zhu J. Towards better analysis of machine learning models: a visual analytics perspective. Visual Informatics. 2017; 1(1):48-56.
[10] Khatavkar V, Velankar M, Kulkarni P. Multi-perspective analysis of news articles using machine learning algorithms. International Journal of Computer Applications.2019.
[11] Celli F, Cumbo F, Weitschek E. Classification of large DNA methylation datasets for identifying cancer drivers. Big Data Research. 2018; 13:21-8.
[12] Khalifa S, Martin P, Young R. Label-aware distributed ensemble learning: a simplified distributed classifier training model for big data. Big Data Research. 2019; 15:1-11.
[13] Genevès P, Calmant T, Layaïda N, Lepelley M, Artemova S, Bosson JL. Scalable machine learning for predicting at-risk profiles upon hospital admission. Big Data Research. 2018; 12:23-34.
[14] Sun N, Sun B, Lin JD, Wu MY. Lossless pruned naive bayes for big data classifications. Big Data Research. 2018; 14:27-36.
[15] McGinnis RS, McGinnis EW, Hruschak J, Lopez-Duran NL, Fitzgerald K, Rosenblum KL, et al. Wearable sensors and machine learning diagnose anxiety and depression in young children. In EMBS international conference on biomedical & health informatics (BHI) 2018 (pp. 410-3). IEEE.
[16] Dumitrescu E, Hue S, Hurlin C, Tokpavi S. Machine learning for credit scoring: improving logistic regression with non linear decision tree effects (Doctoral dissertation). 2018.
[17] Xin M, Wang Y. Research on image classification model based on deep convolution neural network. EURASIP Journal on Image and Video Processing. 2019.
[18] Gupta A. Current research opportunities of image processing and computer vision. Computer Science. 2019; 20(4):387-410.
[19] Bianco S, Cusano C, Piccoli F, Schettini R. Personalized image enhancement using neural spline color transforms. IEEE Transactions on Image Processing. 2020; 29:6223-36.
[20] Liu CL, Shih KT, Huang JW, Chen HH. Light field synthesis by training deep network in the refocused image domain. IEEE Transactions on Image Processing. 2020; 29:6630-40.
[21] Liu S, Thung KH, Lin W, Yap PT, Shen D. Real-time quality assessment of pediatric MRI via semi-supervised deep nonlocal residual neural networks. IEEE Transactions on Image Processing. 2020; 29:7697-706.
[22] Yasarla R, Perazzi F, Patel VM. Deblurring face images using uncertainty guided multi-stream semantic networks. IEEE Transactions on Image Processing. 2020; 29:6251-63.
[23] Mishra A. Metrics to evaluate your machine learning algorithm. Towards Data Science. 2018.

Divya Chauhan received her B.Tech degree in 2012 from Himachal Pradesh University, Shimla, Himachal Pradesh, India. In 2014, she received her M.Tech. degree from HPU, Shimla, Himachal Pradesh, India. She received JRF for three years and is pursuing her Ph.D. degree from HPU, Shimla, India. She is currently appointed as Assistant Professor of Computer Applications in Government College Rampur Bsr., Himachal Pradesh, India.
Email: dvcherish90@gmail.com

K.L. Bansal received his B.Sc. degree, MCA degree and Ph.D. degree in Computer Science and Applications from HPU, Shimla, India in year 1989, 1994 and 2005 respectively. He is currently working as Professor in the Department of Computer Science in HPU, Shimla with 25 years of teaching experience. He has served as the Chairman of the same department for four years. He has guided many M.Phil, M.Tech and Ph.D. students, and has published two books and more than fifty research papers.
Email: kishorilalbansal@yahoo.co.in