

Accurate and Efficient Diagnosis of Brain Tumour Disease using Kernel Support Vector Machine

Mrs. Mary Jacob¹ Mr. Aswin Herbert Sathish²

^{1,2}Assistant Professor

^{1,2}Department of Bachelor of Computer Application

^{1,2}Kristu Jayanti College, Bangalore India

Abstract— Brain tumor often recognized to be a generation of abnormal cells within brain. Brain tumor needs to be diagnosed accurately for the proper decision making about the disease occurrence, so that appropriate treatment on time can be provided to the patients. However analyzing and diagnosing the brain tumor from the imbalanced data set would be more difficult process which needs to be handled very carefully for the accurate brain tumor diagnosis. Physiological data of patients are the primary and vital entities in healthcare big data analytic. Hence, valid raw data must be collected with an efficient manner in a medical environment. In the existing research work, Integrated feature Selection and Ensemble Classification (IFSEC) is introduced which attempts to accurately predict the brain tumor disease. However, handling large volume of data in the single machine would lead to inaccurate and wrong diagnosis of brain tumor disease which needs to be handled with more concern for the accurate output. This is resolved in the proposed research method by introducing novel method namely Accurate Prediction of Brain Tumor Disease from Big Data Framework (APBTD-BDF). To make ease of large volume of data handling process, in the proposed research work, content aware partitioning is introduced which attempts to divide the health care big data into multiple partitions which can be scheduled on different machines for the classification purpose. To improve the classification accuracy in this research low variance filtering approach is introduced for the selection of the more important features. Classification is done by using Kernel Support Vector Machine (KSVM) approach whose performance is tested based 10 fold cross validation technique. The overall research method is implemented in the MATLAB simulation environment from which it can be proved that the proposed research method can effectively handle the large volume of data and it can predict the brain tumor disease more accurately.

Key words: Brain Tumor, Classification, Prediction, Feature Selection, Partitioning, Large Volume of Data

I. INTRODUCTION

Brain tumor is a more critical disease which is found in many persons nowadays [1]. Brain tumor is a generation of unwanted/abnormal tissues over the brain which would lead to brain dead in case of its growth. Brain tumor needs to be analyzed carefully to avoid the dangerous causes which might occur due to uncontrolled growing of abnormal tissues in the brain. However accurate prediction and efficient treatment of brain tumor would be more difficult task [2]. Brain tumors are mostly predicted by combining the Brian tumor MRI image features gathered from the patients and the characteristics of general brain tumor occurrence. It requires more human interaction to accurately predict the brain tumor

presence where the human would take more time for the diagnosis of brain tumor presence [3].

This can be resolved by introducing the automated system which can predict the brain tumor presence automatically by analyzing and learning the presence of brain tumor [4]. The general way of performing brain tumor diagnosis in the automated manner is to gather the multiple MRI image features from the diseased patient which needs to be learned for the identification common factors. And then patients MRI image features would be analyzed for this presence of common factors. If it is found then the brain tumor would be present else the patient doesn't have brain tumor [5]. However accurate recognition of the brain tumor in the automated manner would be more difficult task due to presence of small volume of information about the brain tumor. This is resolved by gathering large volume of information about the brain tumor which would lead to accurate prediction result.

Big data handling is a more difficult task where it consists of large volume of data's which is more complex to handle. General data mining techniques introduced earlier would not perform well in the big data features which required more human intervention and would lead to inefficient handling and inaccurate results [6]. Thus it is required to introduce the new methods that can support and make ease of big data handling, so that accurate prediction of brain tumor disease can be done. Another problem which can be found which handling big data are presence of noise such as unwanted information, tags, labels and so on. This problem also needs to be focused more for the accurate handling and prediction of brain tumor disease presence. In this research work, handling big data and accurate prediction of brain tumor presence is focused to lead to the efficient performance.

In the proposed research work, Brian tumor patients MRI image feature data set is gathered from the multiple sources. Initially, to make ease of handling large volume of data, division of data into multiple modules is done by introducing the contents aware partitioning approach. And then to improve the prediction accuracy and reduce the computation overhead of handling large volume of data, Low variance filtering approach is introduced which can eliminate the unwanted information present in the database and result with only important features. Finally accurate prediction of brain tumor is performed by learning those features that are extracted in the previous phase based on which brain tumor diagnosis would be done. In this work, brain tumor prediction is ensured by using KSVM approach.

The Overall flow of the research work is given as follows: In this section, general introduction about the brain tumor and the ways of prediction is given. In section 2, varying related research works that are proposed by various researchers are discussed in detail. In section 3, overview of

proposed research method along with its merits and demerits are given. In section 4, experimental evaluation of the proposed research methodology is given which is implemented in the mat lab simulation environment. Finally in section 5, overall conclusion of the research work is given.

II. LITERATURE REVIEW

In this section different related research works that are conducted to achieve the better classification and prediction brain tumor disease by various researchers has been discussed in detail. The research methods and its working procedure are mentioned in this section.

Xiao et al [7] introduced a correlation feature based diagnosis of brain tumor is introduced. It performs brain tumor diagnosis by extracting the features from the preprocessed images which would then be classified in order to predict the disease occurrence. Nandagopal and Rajamony [8] introduced the statistical feature based brain tumour recognition which would extract the featured from the MRI images after them into wavelet representation. Here wavelet conversion is done efficiently by using the two level discrete wavelet transform.

Kalbhani et al [9] performed two level classifications namely normal and abnormal by processing the wavelet converted MRI images. Here the classification is done on the feature vector normalized format. Sindhu mol et al [10] attempted to improve the classification accuracy by extracting the featured from the MRI images which would be processed using ICA method before classification.

Navarro et al [11] introduced the novel method for the accurate and efficient feature extraction process which can extract the optimal features from the MRI images more accurately. These methods attempt to ensure the improved classification accuracy by analysing the method more accurately. Saritha et al [12] introduced the brain tumour diagnosis based on wavelet entropy based approach. This method is proved to provide improved classification performance by recognizing it accurately.

Sumitra and Saxena [13] introduced neural network learning based on tumour diagnosis process which can accurately identify the presence of brain tumour based on accurate and optimal extraction of featured from the images. Jayachandran and Dhanasekharan [14] introduced hybrid approach for the accurate diagnosis of brain tumour presence which would perform classification by combining the statistical method and support vector machine.

Nantha Gopal and Sukanesh [15] introduced the automated diagnosis of segmentation and classification of brain tumour presence. This technique attempted to extract both statistical and co-occurrence featured present in the images. Deepa and Devi [16] introduced the back propagation and radial basis diagnosis method for the accurate diagnosis of the brain tumour as cancerous or non cancerous tissues.

Mustara & Suchalatha [17] brain cancer detection and classification approaches for the efficient diagnosis of the brain tumour present in the humans. In this work, gray level co-occurrence matrix (GLCM) method is utilized for the feature extraction process. Jafari and Shafaghi [18] introduced the hybrid approach for the diagnosis of the brain tumour disease by adapting the genetic and support vector machine algorithms.

Rathi and Palani [19] focused on efficient extraction and analysis of the features present in the brain tumour MRI images. This method adapts principal component analysis and the linear discriminate analysis technique for the accurate extraction of the features present in the MRI images. Zarandi et al [20] introduced the type II fuzzy expert system for the accurate diagnosis of the brain tumour disease present in the human disease. This method adapts fuzzy rule base system for the accurate diagnosis. Zhang et al [21] introduced the neural network based MRI image classification which attempts to predict whether the brain tumour is present in the humans or not.

III. ACCURATE AND EFFICIENT PREDICTION OF BRAIN TUMOR DISEASE

Brain tumor recognition plays a more important role in the health care application where the analyzing and predicting the brain tumor would be more difficult task. This problem is focused in this research work which is resolved by introducing the automated system which can predict the brain tumor accurately by using learned information. However training based brain tumor diagnosis requires learning more number of training data with brain tumor to ensure the accurate decision making. However handling of big data and processing them would be more difficult task which needs to be concentrated more for the accurate and efficient decision making process. Thus it is required to introduce the novel techniques that focus on the accurate and efficient detection of the brain tumors. In the proposed research method, efficient handling of big data and predicting more accurate results are focused by introducing the novel framework namely Accurate Prediction of Brain Tumor Disease from Big Data Framework (APBTD-BDF). The steps followed in the proposed research method are listed as follows:

- To improve the classification accuracy in this research low variance filtering approach is introduced for the selection of the more important features
- To make ease of large volume of data handling process, in the proposed research work, content aware partitioning is introduced which attempts to divide the health care big data into multiple partitions which can be scheduled on different machines for the classification purpose
- Classification is done by using KSVM approach whose performance is tested based 10 fold cross validation technique

A. Feature Selection using Low Variance Filtering Approach

Brain tumor big data set gathered from the online web sources would consists of more volume of features among which most of them would not represent the required information. Carrying and handling those unwanted features would lead to reduced classification accuracy and increased computational overhead. Thus it is required to eliminated those noises and unwanted features, so that accurate classification can be done with the processing of required features alone. In this research work, handling and feature reduction is done by using the Low Variance Filter method whose main goal is to eliminate the repeated and unwanted information/ featured present in the gathered brain tumor big data set.

Low variance filtering technique is based on the variance estimation which would calculate the variance between the multiple columns based on which feature reduction would be done. This method would initially calculate the variance present between the multiple columns and the columns with the lower variance level i.e. columns that are similar together would be eliminated. This is measured by fixing the threshold value with which these comparisons would be made. In this research work variance is calculated based on the column values thus the missing values or irregular representation of column values would lead to reduced performance which needs to be tolerated first before applying the dimensionality based variance detection approach. This can be resolved by using the normalization technique which would organize the database values in the regularized manner and the missing values would be replaced accurately. Generally normalization is done in the range between [0, 1] where every column values would be converted into the range of 0 to 1. And then variance between the different columns of database would be calculated. This variance value will be compared with the threshold value, if it is less than the threshold value then it would be found as more similar column with each other and it will be eliminated from the database. This process would be repeated for the every column that is present in the database until completion of comparison between every column. Finally remaining columns would be utilized for the prediction process which would re-normalized again to its original values before performing prediction task. The flow of Applying Low variance filtering approach on the Brain tumor big data set is shown in the following figure 1.

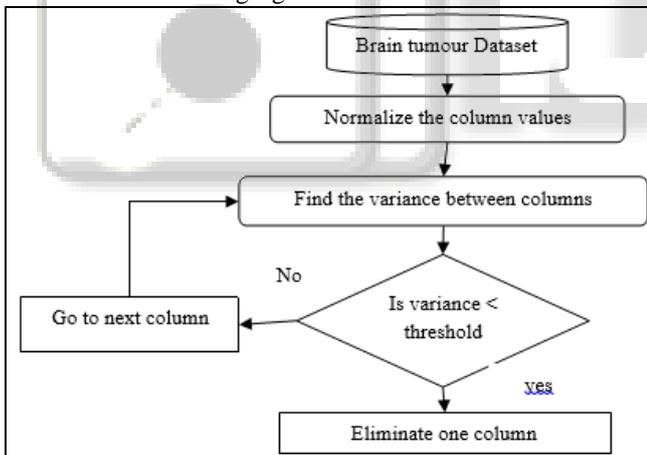


Fig. 1: Low variance based feature selection

B. Content Aware Partitioning Method

After feature elimination and selection, remaining features need to be divided into multiple partitions, as those are big in size which cannot be handled in the single machine. Handling volume of data in single processor with limited configuration would lead to inaccurate decision making due to more computation overhead. And also this technique would lead to reduced system performance where the decision making cannot be done on right time. This problem can be avoided by dividing the data into multiple partitions which attempts to reduce the burden of handling large volume of data in the single machine. In the proposed research work, this problem is resolved by introducing the content aware partitioning method which would divide the entire dataset into multiple

sub data in terms of column separation. This method performs partitioning by considering the knowledge of the contents of the different columns present in the data set.

Content aware partitioning method main goal is to divide the entire data set into multiple partitions in terms of its columns and the size of the columns. This process would divide the entire column into multiple subsets of columns based on knowledge gathered from the multiple tuple and combining those knowledge levels. This would be analyzing the tuple present in the database and tuple with uniform interval would be combined together as partitioning columns. And then corresponding blocks to execute those partitions is allocated by adapting the technique called the fair allocation block placement strategy. Finally Low variance filtering would be applied on the divided columns once again to avoid the irrelevant or more similar columns present in the sub partitions.

In this research work brain tumor data set is considered in which irrelevant columns such as race, religion and so on which are irrelevant to the disease prediction would be eliminated from the database. After reduction of number of features it will be given as input to the content based partitioning approach where the entire data set would be divided into multiple columns such as cancerous disease data as one partition and the non-cancerous disease as another partition. These partitioning would be assigned to the multiple blocks for the proper decision making where the prediction methodology would be applied. The work flow of this methodology is given in figure 2 as follows:

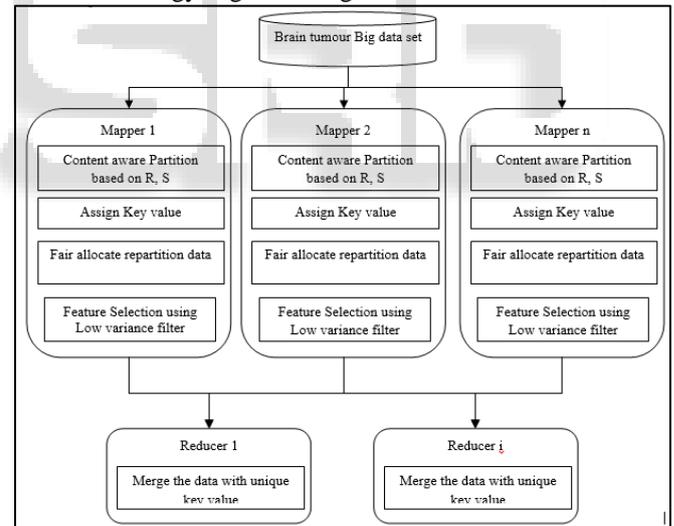


Fig. 2: Overall flow of the research work

C. Classification using Kernel Support Vector Machine

After partitioning the data set, prediction process is carried over on the different partitions present in the different blocks. In this research work KSVM is applied to perform the prediction process whose main goal is to learn the feature behaviour of both cancerous and non-cancerous effects of brain tumor disease. Traditional classification algorithms which cannot easily handle huge amounts of data. So the prediction is performed by using Kernel Support Vector Machine (KSVM).

1) Kernel Support Vector Machine (SVM) classifier

The classification algorithm starts with a review of Support Vector Machines (SVMs) for classification of . Given labeled training data of the form $\{(y_i, x_i)\}_{i=1}^N$ with $y_i \in \{-1, +1\}$, $x_i \in$

R^n , use a CSVM formulation [32]. For a kernel on data points, $k(x, z) : R^n \times R^n \rightarrow R$, that is the inner product, $\Phi(x) \cdot \Phi(z)$, in an unrealized, possibly high dimensional, feature space, the algorithm finds a hyperplane which best separates the data by minimizing:

$$\tau(w, \xi) = \frac{1}{2} \|w\|^2 + C \sum_{i=1}^N \xi_i \quad (1)$$

Subject to $y_i((w \cdot x_i) + b) \geq 1 - \xi_i$ and $\xi_i \geq 0$, where $C > 0$ is the tradeoff between regularization and constraint violation. In the dual formulation maximize:

$$W(\alpha) = \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{ij} \alpha_i \alpha_j y_i y_j k(X_i, X_j) \quad (2)$$

Subject to $0 \leq \alpha_i \leq C$ and $\sum_{ij} \alpha_i y_j = 0$

The decision function is $\text{sign}(h(x))$, where

$$h(x) = \sum_{i=1}^m \alpha_i y_i k(x, X_i) + b \quad (3)$$

For clarity, in a slight abuse of notation the features $x_i, i \in \{1, 2, \dots, m\}$ will be referred to as support vectors. Thus, in general m kernel computations are needed to classify a point with a kernelized SVM and all m support vector must be stored. For linear kernels we can do better because, $k(x, z) = (x, z)$, so $h(x)$ can be written as $h(x) = (w, x) + b$, where $w = \sum_{i=1}^m \alpha_i y_i x_i$. It is possible to speed up classification for IKSVMs. For feature vectors $x, z \in R_+^n$, the intersection kernel is $k(x, z)$:

$$k(x, z) = \sum_{i=1}^n \min(x(i), z(i)) \quad (4)$$

and classification is based on evaluating: (5)

$$\begin{aligned} h(x) &= \sum_{i=1}^m \alpha_i y_i k(x, x_i) + b \\ &= \sum_{i=1}^m \alpha_i y_i \left(\sum_{j=1}^n \min(x(j), x_i(j)) \right) + b \end{aligned} \quad (6)$$

The non-linearity of \min prevents us from 'collapsing' in a similar manner to linear kernels. This is in general true for any nonlinear kernel including radial basis functions, polynomial kernels, etc.

IV. EXPERIMENTAL RESULTS

The experiments are conducted in the matlab simulation environment where the performance evaluation between the proposed method namely Accurate Prediction of Brain Tumor Disease from Big Data Framework (APBTD-BDF) and the existing methods namely Integrated feature Selection and Ensemble Classification (IFSEC), Artificial Neural Network based Prediction (ANN), Map reduce Based Partitioning (MRP) is done. The performance measured that are considered in this research work for the comparison evaluation are, "Accuracy, Precision, Recall, F-Measure, Mean absolute error and computation time". The performance evaluation are shown and explained in the following sub sections.

A. Accuracy

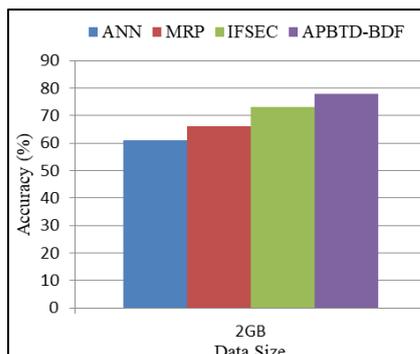


Fig. 3: Accuracy

Accuracy is defined as the proportion of true positives and true negatives among the total number of results obtained. Accuracy is evaluated as,

$$\text{Accuracy} = \frac{(\text{True positive} + \text{True negative})}{(\text{True positive} + \text{True negative} + \text{False positive} + \text{False negative})}$$

From the above graph mentioned in figure 3, it is proved that the proposed method namely APBTD-BDF can provide better performance result than the existing research methodologies. From the analytical evaluation of the results, it can be proved that the proposed methodology APBTD-BDF shows 7% better performance than IFSRC method, 18% better performance than MRP method and 28% better performance than the ANN method.

B. Precision

Precision value is evaluated according to the relevant information at true positive prediction, false positive.

$$\text{Precision} = \frac{\text{True Positive}}{(\text{True Positive} + \text{False Positive})}$$

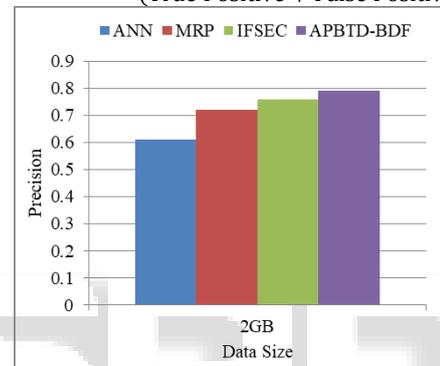


Fig. 4: Precision

From the above graph it is proved that the proposed method namely APBTD-BDF can provide better performance result than the existing research methodologies in terms of precision. From the analytical evaluation of the results, it can be proved that the proposed methodology APBTD-BDF shows 4% better performance than IFSRC shows 10% better performance than MRP method and 30% better performance than the ANN method.

C. Recall

Recall value is evaluated according to the retrieval of information at true positive prediction, false negative.

$$\text{Recall} = \frac{\text{True Positive}}{(\text{True positive} + \text{False negative})}$$

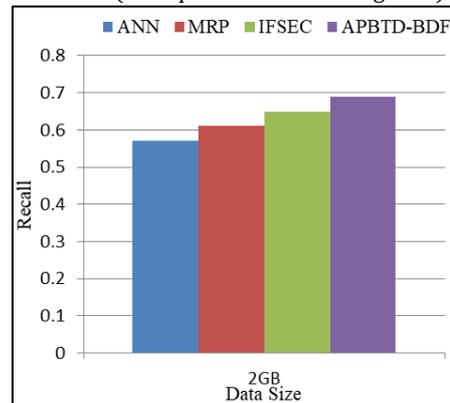


Fig. 5: Recall

From the above graph it is proved that the proposed method namely APBTD-BDF can provide better

performance result than the existing research methodologies. From the analytical evaluation of the results, it can be proved that the proposed methodology APBTD-BDF shows 6% better performance than IFSRC method, 13% better performance than MRP method and 21% better performance than the ANN method.

D. F-Measure

The F-Measure computes some average of the information retrieval precision and recall metrics

$$F - Measure = \frac{2 * precision * recall}{precision + recall}$$

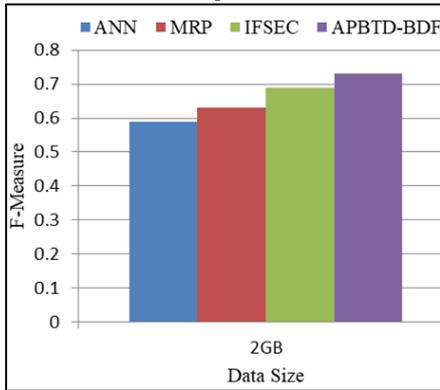


Fig. 6: F-measure

From the above graph it is proved that the proposed method namely APBTD-BDF can provide better performance result than the existing research methodologies. From the analytical evaluation of the results, it can be proved that the proposed methodology APBTD-BDF shows 6% better performance than IFSRC method, 16% better performance than MRP method and 24% better performance than the ANN method.

E. Mean Absolute Error

The mean absolute error (MAE) is a quantity used to measure how close forecasts or predictions are to the eventual outcomes. The mean absolute error is given by

$$MAE = \frac{1}{n} \sum_{i=1}^n |f_i - y_i| = \frac{1}{n} \sum_{i=1}^n |e_i|$$

As the name suggests, the mean absolute error is an average of the absolute errors $|e_i| = |f_i - y_i|$, where f_i is the prediction and y_i the true value. Note that alternative formulations may include relative frequencies as weight factors.

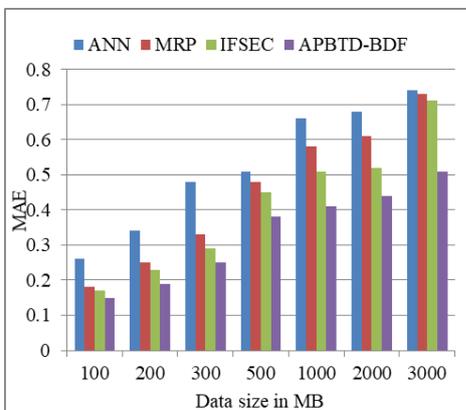


Fig. 7: MAE comparison

From the above graph it is proved that the proposed method namely APBTD-BDF can provide better

performance result than the existing research methodologies. From the analytical evaluation of the results, it can be proved that the proposed methodology APBTD-BDF shows 19% better performance than IFSRC method, 26% better performance than MRP method and 37% better performance than the ANN method.

F. Computation Time

Computation time is defined as the total time consumed for the filtering and tagging process.

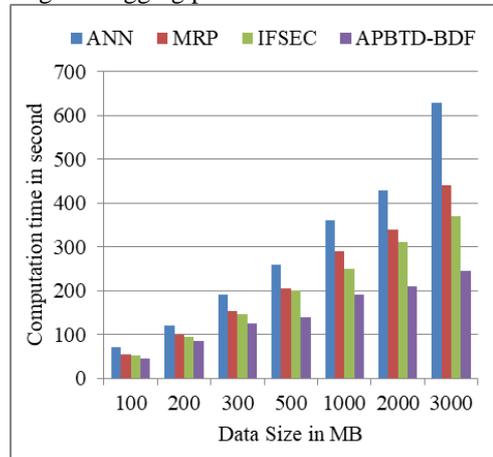


Fig. 8: Computation time

From the above graph it is proved that the proposed method namely APBTD-BDF can provide better performance result than the existing research methodologies. From the analytical evaluation of the results, it can be proved that the proposed methodology APBTD-BDF shows 27% better performance than IFSRC method, 34% better performance than MRP method and 50% better performance than the ANN method.

V. CONCLUSION

Brain tumor recognition from the large volume of training data plays more important role in the real world environment which needs to be focused well for the accurate prediction result. This is focused in the proposed research method by introducing novel method namely Accurate Prediction of Brain Tumor Disease from Big Data Framework (APBTD-BDF). To make ease of large volume of data handling process, in the proposed research work, content aware partitioning is introduced which attempts to divide the health care big data into multiple partitions which can be scheduled on different machines for the classification purpose. To improve the classification accuracy in this research low variance filtering approach is introduced for the selection of the more important features. Classification is done by using Kernel Support Vector Machine (KSVM) approach whose performance is tested based 10 fold cross validation technique. The overall research method is implemented in the MATLAB simulation environment from which it can be proved that the proposed research method can effectively handle the large volume of data and it can predict the brain tumor disease more accurately.

REFERENCES

[1] Xiao, K., Liang, A. L., Guan, H. B., & Hassanien, A. E., "Extraction and Application of Deformation Based Feature in Medical Images", Volume:120, PP:177-184, ELSEVIER Neuro computing, 2013

- [2] Nanthagopal, A. P., & Sukanesh, R., "Wavelet statistical feature based segmentation and classification of brain computed tomography images", Volume:7, No:1, PP:25-32, IET Image Process, 2013
- [3] Kalbkhani, H., Shayesteh, M. G., & Zali-Vargahan, B., "Robust algorithm for Brain Magnetic Resonance Image Classification based on GARCH variances Series", Volume:8, No:6, PP:909-919, ELSEVIER Biomedical Signal Processing and Control, 2013
- [4] Sindhumol, S., Kumar, A., & Balakrishnan, K., "Spectral clustering independent component analysis for tissue classification from brain MRI". Volume:8, No:6, PP:667-674, ELSEVIER, Biomedical signal processing and control, 2013
- [5] González-Navarro, F. F., Belanche-Muñoz, L. A., Romero, E., Vellido, A., Julià-Sapé, M., & Arús, C., "Feature and model selection with discriminatory visualization for diagnostic classification of brain tumour", Volume:73, No:4, PP: 622-632, ELSEVIER, computerized Medical Imaging and Graphics, 2010
- [6] Saritha, M., Joseph, K. P., & Mathew, A. T., "Classification of MRI brain images using combined wavelet entropy based spider web plot and probabilistic neural network", Volume:34, No:16, PP:2151-215, ELSEVIER, pattern Recognition letters, 2013
- [7] Sumitra, N., & Saxena, R. K., "Brain Tumor Classification Using Back Propagation Neural Network", Volume:5, No:2, PP:45-50, International Journal of Image, Graphics and Signal Processing, 2013
- [8] Jayachandran .A & R.Dhanasekaran , "Brain tumor Detection and Classification of MRI Using Texture Feature and Fuzzy SVM Classifiers", Volume:6, No:12, PP:2264-2269, Research Journal of Applied Sciences, Engineering and Tech, 2013
- [9] Nanthagopal, A. P., & Rajamony, R. S., "Classification of Benign and Malignant Brain Tumour CT images using wavelet texture parameters and Neural Network Classifiers", Volume:16, No:1, PP:19-28, Publication: Journal of visualization society, 2013
- [10] González-Navarro, F. F., Belanche-Muñoz, L. A., Romero, E., Vellido, A., Julià-Sapé, M., & Arús, C., "Feature and model selection with discriminatory visualization for diagnostic classification of brain tumour", Volume:73, No:4, PP:622-632, Neuro Computing, 2010
- [11] Jain, S., "Brain cancer classification using GLCM based feature extraction in artificial neural network", Volume:4, NO:7, PP:966-970, International Journal of Computer Science and Engineering Technology, 2013
- [12] Jafari, M., & Shafaghi, R., "A Hybrid Approach for Automatic Tumour Detection of Brain MRI Using Support Vector Machine And Genetic Algorithm", No:3, PP:1-8, Global Journal of Science, Engineering and Technology(ISSN:2332-2441), 2012
- [13] Gladis V.P, Pushpa Rathi and S.Palani, "Linear Discriminate Analysis For Brain Tumour Classification Using Feature Selection", Volume:5, No:5, International Journal Of communication and Engineering, 2012
- [14] Zarandi, M. F., Zarinbal, M., & Izadi, M., "Systematic Image processing for diagnosis Brain Tumour: A type II expert system approach", Volume:11, No:1, PP:285-294, ELSEVIER applied soft computing, 2011
- [15] Zhang, Y., Dong, Z., Wu, L., & Wang, S., "A Hybrid Method for MRI Brain Image Classification", Volume:38, No:8, PP:10049-10053, ELSEVIER Expert System with Application, 2011
- [16] Singh, S. K., Hawkins, C., Clarke, I. D., Squire, J. A., Bayani, J., Hide, T., ... & Dirks, P. B., "Identification of human brain tumor initiating cells", Volume:432, No:7015, PP:396-401, Nature, 2004
- [17] Vescovi, A. L., Galli, R., & Reynolds, B. A., "Brain tumor stem cells", Volume:6, No:6, PP:425-436 Nature Reviews Cancer, 2006
- [18] Interphone Study Group., "Brain tumor risk in relation to mobile telephone use: results of the INTERPHONE international case-control study", International journal of epidemiology, dyq079
- [19] Grant, R., "Overview: brain tumor diagnosis and management/Royal College of Physicians guidelines", Volume:75, No:2, PP:ii18-ii23, Journal of Neurology, Neurosurgery & Psychiatry, 2004
- [20] Sinkus, R., Lorenzen, J., Schrader, D., Lorenzen, M., Dargatz, M., & Holz, D., "High-resolution tensor MR elastography for breast tumor detection", Volume:45, No:6, PP:1649, Physics in medicine and biology, 2000
- [21] Srinivasa, S., & Bhatnagar, V., "Big data analytics", PP:24-26, In Proceedings of the First International Conference on Big Data Analytics BDA, 2012