Abstract

Objectives: To detect brain cancer in obese and non-obese patients by applying datamining and classification techniques. These techniques will be beneficial for both patients and doctors.

Methods/Statistical Analysis: Data mining is a broader term for a variety of data analysis techniques. These techniques will be applied to extract meaningful knowledge from a large and noisy database. These techniques have special ability to adapt to the local characteristics of the data. Different data mining techniques like data preprocessing, decision tree C5.0 algorithm are used for discovering useful and meaningful information from medical data, like for brain cancer prediction in obese and non-obese patients.

Findings: From the last few years, the Data of Medical Science become vast rapidly. There is a need to extract meaningful knowledge from the large data set by applying data mining techniques. Cancer is one of the major diseases and brain cancer is one of the major types of cancer which become major cause of death in men as well as in women. Brain cancer is also known as brain tumor and it is becoming the cause of death due to brain cancer in Pakistan and Worldwide. Obesity is also one of the major factors of brain tumor. Brain cancer starts when cells of brain become abnormal and grow out of control. Excess body fat changes the level of hormones. These hormones release chemicals in immune cells that cause long lasting inflammation and this inflammation in cells can raise the risk of cancer. In this paper, we shall discover the relationship of obesity with brain cancer and how obesity is also a major cause of brain tumor.

Application/Improvements: The results which we concluded from this method can be used as supporting or assistant tool for neuro-oncologist for identification and diagnosis of brain tumors.

Keywords: Carcinogens, Data Mining (DM), Meningioma, Neuro-oncologist, Obesity, Pre-Processing

1. Introduction

Data mining is a set of techniques to analyze data which came from various sources and discover the knowledge from it to use specifically. Such valuable knowledge can be used to solve problems like health care, causes of diseases and their cure. In current practice, health care industry gather large volume of data related patients, diseases, diagnosis and cure. The different methods of data-mining facilitate a set of rules to find useful patterns and meaningful information from data. Provision of quality of service to patient is a huge challenge for medical health care industry. Quality of services to patients signifies early detection or diagnosing diseases correctly and provides effective treatment to suffered patients. Improper diagnosis can lead to devastating results which are not acceptable to anyone. For example brain cancer which is a major cause of death, can be curable only if detects in early stages of patients.

1.1 Background

World Health Organization (WHO) had arranged a survey about brain cancer in which disclosed that brain cancer is a major cause of death around the world. Medically brain
cancer occurs due to abnormality in life cycle of cells production in brain area. These abnormalities of normal cells which results of the interaction between a person's genetic factors (also called Primary brain tumor), age factor, already affected with some other cancer type (also called Secondary brain tumor) and carcinogens (physical/chemical/biological). Brain cancer may happened due to unhealthy diet with low fruit and vegetable intake also obesity in humans. Obesity is a cause of deaths become doubled since 1989, killed more people than underweight in most of the countries. Obesity or overweight impair health and become harmful in producing hormones which harmfully affect the way of working of brain cells. Due to these harmful effects, the risk of brain cancer can also increase.

Diagnosis of brain cancer is much complex and essential job that need to be performed correctly and proficiently. There are series of different types of medical test results are necessary for detecting brain cancer. Practically, classification techniques are used in the field of medical sciences. Most data mining methods from classification category are used as the applied prediction techniques. Therefore a predictive detection system is proposed that can take benefits from medical database and DSS. Through using that system, we will be able to categorize the patients into different categories and figure out brain cancer into obese/non obese patients.

The core purpose of this study is to apply data mining technique for extracting hidden patterns, which are important in early detection of brain cancer in obese/non obese patients. For this we are identifying key attributes which are more relevant in relation for brain cancer detection, applying decision tree for predicting brain cancer.

2. Relationship between Brain Cancer and Obesity

It is very important to know that brain is a most important body part. It works as the core part of the nervous system. It is most complicated structure organ of a human body. Brain has two types of functions. Firstly, it controls the rest of the human body by producing patterns for muscle activity and secondly by driving the secretion of chemicals in whole human body. These body chemicals called hormones. Brain cancer/tumor occurs in brain area. In human body usually cells become old or get damage and they expire so new cells born and take the place of died cells. But due to some reasons sometimes the life cycle of cell goes wrong and body starts to produce new cells, when body does not need new cells. The production of additional cells can produce an extra mass of tissue in body. This extra mass of tissue is called tumor. On the basis of cells, human body has two types of brain tumor.

• Benign tumor (do not contain cancer cells).
• Malignant tumor (contain cancer cells).

Malignant tumor is the main cause of different brain cancers which effect on different sides of brain. The symptoms of brain cancer depend on size and location within the central nervous system. The most common symptoms of brain cancer are headache, fatigue, drowsiness, seizures (myoclonic and tonic-clonic), difficulty in fine motor skills, loss of vision, change in sensory, inability to look upward, changes in lactation and menstrual periods and abnormal growth in hands or feet.

It is very depressing that the actual main reason of brain tumor is unknown. But we can highlight different risk factors to understand to understand brain cancer. The most common risk factors are: genetic factor, previous cancer, viral infection, age factor, exposure to radiations and being overweight or obese.

Obesity increases the risk of brain cancer like meningioma (a type of brain tumor). Obesity defined as a body with an excess weight or overweight. An obese body has Body Mass Index (BMI) of 30 or higher than 30. Excess body fat can changes the level of secretion of hormones in body. These hormones produce proteins that released into the blood and this blood carried these hormones around the whole human body. These changes in hormones can affect many parts of the body. It can also raise the threat of different types of cancer in human body including brain cancer. These fat abnormal cells can also affect other cells of body including immune cells of brain tissues. And these immune cells produce chemicals that produce harmful inflammation in brain cells. This inflammation in cells can last for long time. It can raise the risk of brain cancer.

There has been already so many researchers done that have main emphasis on identification, prediction and detection of brain tumor. They have used different techniques and methods of data mining for identification and for prediction brain tumor in patients.

• An image mining technique of classification for brain tumor detection by using pruned association rule with MARI algorithm was proposed by. It uses association rules for classifying the CT scan images of brain into
categories. And extract the low level features and high level knowledge from CT scan brain images for specialists. And then combined all the low and high level features for improving the accuracy and assisting the physicians.

- A method of classification is used for diagnosing brain tumor in patients, using MRI images by\(^3\). They have used preprocessing, feature extraction, association rule mining and classification. The efficiency of image mining methods will enhanced by using their proposed model. Decision Tress and Naïve Bayesian Classification algorithm was used for comparing accuracy.

- A model was proposed for brain tumor detection from MRI images by\(^4\). They have used segmentation in image classification and analysis. They used automatic segmentation of brain MRI to correctly segment the image in different segments in a small interval of time. And then classify them according to categories of brain tumor.

- A method for early detection and prevention of cancer using data mining techniques was proposed by\(^5\). They have used a method by combining K-Means clustering algorithm for separating cancer and non-cancer patients and decision trees classification algorithm for classifying the patients according to cancer type.

- A model was proposed by\(^6\)-which use CT images in classification and segmentation for classifying brain tumor. For extracting brain tumor from CT images, they used Support Vector Machine Classifier (SVMC) segmentation algorithm. Then they used classification for classifying the brain tumor. This model helps in segmentation and classification of brain tumor for detection in patients.

- A model which use segmentation for segmenting the brain MRI images results and detect the brain tumor on the basis on multi parameters related specific tumor and then use classification for classifying the tumor was proposed by\(^7\). This process will helps in detection of brain tumor from MRI images of patient.

3. **Data Mining Technique used for Prediction**

In this paper, we are using decision tree mining technique for designing prediction system for detection brain cancer in obese/non obese patients. Decision tree data mining classification technique is used to analyze the dataset.

### 3.1 Decision Tree

The decision tree technique is commonly used for classification problems. There are various techniques of decision tree - g. CART, ID3, C4.5, CHAID, J48 and C5.0. In this paper we are using C5.0 algorithm to build a decision tree. It is improved version of C4.5. This technique is more effective in performance and accuracy wise than C4.5. The main task of this algorithm is to predict a class from the values of attributes. Its feature selection ability reduced its error rate. It identifies noisy and missing data and also solves the problem of error pruning. It can predict only categorical targets. C5.0 also offers the high boosting method to increase the accuracy of classification. So in this paper, we are using C5.0 algorithm for splitting the data according to selected attributes and that provide the maximum information gain and helps in deriving new attributes for getting more accurate results.

4. **Data Preprocessing**

Incomplete, missing, noisy and inconsistent data are common place issues of large real world databases. For handling missing values, remove noise, to handle outliers and correct inconsistencies in database, we used data preprocessing techniques.

### 4.1 Data Collection

11848 Patient’s data was obtained from SKMCH&RC, Lahore, Pakistan database which were suffered in different cancers. But we have used only 527 brain cancer patient records in our research. There are 343 male patients whose age from 0-85 years old and 184 female patients whose age from 0-87 years old.

### 4.2 Data Format and Analysis

Data formatting is a process in which we convert our different format distributed data into single format for using data easily. This process need different conversions for making data more descriptive and summarize. Then we analyze the number of attributes from database and extract only important and interesting attributes for enhancing the effectiveness of system.
Early Detection of Brain Cancer in Obese and Non-Obese Patients by using Data Mining Techniques

4.3 Data Cleansing and Integration
The phase in which, we handle noisy data, missing values, handling outliers and handle inconsistent data for making data more consistent. This phase was performed after addressing requirements of tool which we selected for this preprocessing phase. Then we integrate all multiple data sources into single data source.

4.4 Data Selection and Transformation
In this phase, the most relevant data to analysis is retrieved from records. For retrieving relevant data we use feature selection technique. After this phase, we transformed the data into appropriate form for analysis. For this we used discretization, data reduction technique and attribute selection process.

5. Knowledge from Data
Now many hospitals were using health information system; these systems contain huge amount of data this data can be used to find out hidden useful medical information for making intelligent medical diagnosis system.

Currently a neuro-oncologist, a doctor is present in medical center for brain tumor patients. He/she specialized in diagnosing and treating brain tumors in patients. For the identification of location of tumor, doctor uses symptoms of patient’s body health as a clue. In addition to a patient’s detailed medical history and entire body examination and medical test (e.g., MRI test, CT Scan test, X-Ray test, Cerebral Angiogram test, PET Scan test, Electroencephalography test, Neurological test, Vision and hearing test) results may help the doctor in finding and diagnosis brain tumor.

Our main research purpose is to make an intelligent brain cancer detection technique that detects brain cancer in obese/non obese patients using historical brain cancer database. To develop this system, we are using database of SKMCH&RC, Lahore, Pakistan. The total database consists of 11848 cancer patient’s record from 2014 to 2015. There are 527 records of brain cancer patients. We only use these 527 records for detecting our required results.

Firstly selected dataset contained some fields with missing values in some records. These records with missing values were identified and changed them with most suitable values using method of replace missing values by taking mean or mode of all records. These data handling is done by using data preprocessing technique.

The variables in the data set are consists of two types of attributes:
- Key attributes.
- Input Attributes.

These attributes are listed below.
- Key Attribute
  - MRNO: Patient’s Identification Number
- Input Attributes
  - BMI, Obesity and Obesity to get more appropriate results for detection brain tumor in obese/non obese patients, as shown in Table 3. During data preprocessing, we find these few statistics of brain cancer patients in our data.

In Table 4, the gender wise cancer counts, this table shows rate of cancer is greater in females as compared to males. In male and female cancer ratio is 48.30% and 51.70% respectively.

Table 5 shows gender wise brain cancer counts, as table shows brain cancer is greater in male as compared to female. The occurrence of brain cancer in male and female is higher between the ages of 20 to 60.

Table 6 and Figure 1 shows the gender wise brain cancer counts, as figure shows brain cancer is greater in male as compared to female. The occurrence of brain cancer in male and female is higher between the ages of 20 to 60.

Table 7 shows that mostly brain cancer patients are laying in secondary brain tumor category which means they are already affected with some other cancer type.

![Figure 1. Distribution of brain cancer patients (year wise).](image-url)
As Table 8 shows the gender wise distribution of brain cancer patients which are affected with different types of brain tumor. According to table results, mostly males are affected with Glioblastoma and females with Meningioma.

The distribution of obese/non-obese brain cancer patients is shown in Table 9. As table shows brain cancer is greater in non-obese as compared to obese. But the ration of obese brain cancer in females is higher than males.

![Figure 2](image)

**Figure 2.** Distribution of obese brain cancer patients (brain tumor category wise w.r.t gender).

Figure 2 shows the distribution of obese brain cancer patients according to brain tumor categories (primary brain tumor, secondary brain tumor) w.r.t gender. As figure shows primary brain cancer is greater in obese females as compared to males which means obesity can be one major cause in causing cancer to these cancer female patients.

![Figure 3](image)

**Figure 3.** Distribution of obese brain cancer patients according to different types of brain tumor.

Figure 3 shows the distribution of obese brain cancer patients according to different types of brain tumor, as figure shows that the rate of meningioma is higher in female obese brain tumor patients. Total 22.05% brain cancers patients are obese in overall 527 brain cancer patients. And 35.34% and 64.66% brain cancer occurs in

| Sl.No | Attributes | Description | Values |
|-------|------------|-------------|--------|
| 1     | MRNO       | Patient’s Identification Number | Numerical |
| 2     | Age        | Age in years | Numerical |
| 3     | Sex        | Sex | Nominal (1=Male, 2=Female) |
| 4     | Height     | Height in ft/in | Numerical |
| 5     | Weight     | Weight in kg | Numerical |
| 6     | Diagnosis  | Diagnosis (Does patient has some cancer already) | Nominal (0=Not Present, 1= Present) |
| 7     | Family_His | Family History of Brain Tumor | Nominal (0=Not Present, 1= Present) |
| 8     | Expo_Radi  | Exposure of radiations in home/work place | Nominal (0=Not Present, 1= Present) |
| 9     | Seizures   | Seizures (Myoclonic, Generalized tonic-clonic) | Nominal (0=No, 1= Yes) |
| 10    | Loss_Balance | Loss of balance/difficulty in fine motor skills | Nominal (0=No, 1= Yes) |
| 11    | Loss_Vision | Loss of vision (Partial/Complete) | Nominal (0=No, 1= Yes) |
| 12    | Chg_Senses | Changes in sensory (Sensation, Vision, Smell, Hearing) | Nominal (0=No, 1= Yes) |
| 13    | Diffi_Swal | Difficulty in swallowing | Nominal (0=No, 1= Yes) |
| 14    | Abnorm_Lacta | Abnormality in lactation period | Nominal (0=No, 1= Yes) |
| 15    | Inab_Look_Up | Inability to look upward | Nominal (0=No, 1= Yes) |
| 16    | Abnorm_Mens | Abnormality in menstrual periods | Nominal (0=No, 1= Yes) |
| 17    | Abnorm_Grow | Abnormal growth in hands/feet | Nominal (0=No, 1= Yes) |
Early Detection of Brain Cancer in Obese and Non-Obese Patients by using Data Mining Techniques

Table 2. Input attributes (derive from medical results) with description

| Sl.No | Attributes | Description                        | Values                                                                 |
|-------|------------|------------------------------------|------------------------------------------------------------------------|
| 18    | MRI_R      | MRI Results                         | Nominal (0=Positive, 1=Negative) If there is cancer then results show negative otherwise results will be positive |
| 19    | CT_Scan_R  | CT Scan Results                     | Nominal (0=Positive, 1=Negative) If there is cancer then results show negative otherwise results will be positive |
| 20    | XRay_R     | X-Ray Results                       | Nominal (0=Positive, 1=Negative) If there is cancer then results show negative otherwise results will be positive |
| 21    | CAG_R      | Cerebral Angiogram Results          | Nominal (0=Positive, 1=Negative) If there is cancer then results show negative otherwise results will be positive |
| 22    | PET_Scan_R | PET Scan Results                    | Nominal (0=Positive, 1=Negative) If there is cancer then results show negative otherwise results will be positive |
| 23    | ERG_R      | Electroencephalography Results      | Nominal (0=Positive, 1=Negative) If there is cancer then results show negative otherwise results will be positive |
| 24    | Neuro_R    | Neurological Results                | Nominal (0=Positive, 1=Negative) If there is cancer then results show negative otherwise results will be positive |
| 25    | VisHer_R   | Vision and Hearing Results          | Nominal (0=Positive, 1=Negative) If there is cancer then results show negative otherwise results will be positive |

Table 3. Newly added input attributes

| Sl.No | Attributes | Description                        | Values                                                                 |
|-------|------------|------------------------------------|------------------------------------------------------------------------|
| 26    | BMI        | BMI                                | Numerical (it is calculated by BMI=(weight in kg/height in meters)     |
| 27    | Obese      | Obesity                            | Nominal (0=Yes, 1=No) If Obesity level is occurring in Class 1 or Class 2 then body is obese |
| 28    | Obes_Lel   | Obesity Level                      | Nominal (Class 1 = C1, Class 2 = C2, Class 3 = C3) Divide the BMI into different classes according to different ranges of BMI. Class 1 : BMI >= 30.0 then it is obese body Class 2 : BMI ranges 25.0 to 29.0 then it is overweight Class 3 : BMI < 25.0 then it is normal weight body |

Table 4. Distribution of overall cancer patients (gender wise)

| Total Overall Cancer Patients from 2015 to 2016 | Males (%) | Females (%) |
|------------------------------------------------|-----------|-------------|
| 11848                                          | 5722 (48.30%) | 6126 (51.70%) |

Table 5. Distribution of brain cancer patients (gender wise)

| Total Brain Cancer Patients from 2015 to 2016 | Males (%) | Females (%) |
|---------------------------------------------|-----------|-------------|
| 527                                          | 343 (65.09%) | 184 (34.91%) |

“Primary Brain Cancer” and “Secondary Brain Cancer” respectively.

6. Experimentation

Firstly we applied data preprocessing techniques on our dataset for removing missing, irrelevant, noisy values from dataset. These dataset handling is done by using different techniques of data preprocessing. After applying data preprocessing methods on data, data mining techniques of classification are applied. In this research, Decision Trees (C5.0 Algorithm) were applied on data. For applying C5.0 algorithm, we used a tool Statistical Packages for the Social Sciences (SPSS) Clementine for experimentation. In our selected dataset, there are total 527 records of brain cancer patients. These 527 records are divided into two
Table 6. Distribution of brain cancer patients

| Age Groups | Total Patients | Male | Female |
|------------|----------------|------|--------|
| Age < 18 years | 19             | 10   | 9      |
| Age > 18 years | 508            | 333  | 175    |

Table 7. Distribution of brain cancer patients

| Brain Cancer Category | Total Patients | Male | Female |
|-----------------------|----------------|------|--------|
| Primary Brain Tumor   | 140            | 119  | 47     |
| Secondary Brain Tumor | 387            | 224  | 137    |

Table 8. Distribution of brain cancer patients

| Brain Tumor Type     | Males | Females | Total |
|----------------------|-------|---------|-------|
| Glioblastoma         | 110   | 52      | 162   |
| Glioma               | 83    | 43      | 126   |
| Astrocytoma          | 57    | 29      | 86    |
| Oligadendroglioma    | 30    | 14      | 44    |
| Medulloblastoma      | 24    | 0       | 24    |
| Ependymoma           | 25    | 0       | 25    |
| Meningioma           | 14    | 46      | 60    |

Table 9. Distribution of obese brain cancer patients

| Obese Brain Cancer Patients | Total Patients | Male | Female |
|-----------------------------|----------------|------|--------|
| Obese                       | 116            | 42   | 74     |
| Non Obese                   | 411            | 247  | 164    |

Table 10. Results of analysis node

| Results by Analysis node | Total Obese Brain Cancer Patients | Percentage |
|--------------------------|----------------------------------|------------|
| Correct                  | 115                              | 99.5%      |
| Wrong                    | 1                                | 0.5%       |

data sets: training dataset, testing dataset. There is one dataset of 303 records which is used as training and other dataset of 270 records is used for testing.

Figure 4 shows the model of decision trees (C5 algorithm) which applied on our dataset using SPSS Clementine tool. We pass source data into tool by “Source Node” and then derive some new fields for saving the results which we calculate by using our source data. Statistics of results can show using “Histogram Node” which tool derive from “Derive Node”. The metadata/Type Node gives a facility to point out the types of fields that we are using and how these fields are used to predict the outcomes. Then C5 Algorithm node attached to estimate the model. An analysis node is used to compute the accuracy rate of classification technique. An analysis node output shows that with this dataset, the model correctly predicted the rate of obese brain cancer patients. The results of analysis node of decision trees (C5 algorithm) which applied on our dataset using SPSS Clementine are shown in Table 10. It shows the rate of accuracy of classification.

7. Discussion

Obese body is most suitable environment for various diseases and it can become main cause of different cancers including brain cancer. Brain cancer is a dangerous disease but it can be cured by detecting it in early stage. A person life can be saved by detection of brain cancer in early stage and by prevention from spreading in other parts of body. Early detection can help in proper treatment of brain cancer patient. And it also can save patient from death and can increase its chances to live few more. The data mining techniques can play important role in brain cancer detection. Because by these techniques, the probability of brain cancer can be calculate by using results of different medical tests and patient medical history.

8. Conclusion

The main purpose of our work is to design a predictive model for detection the presence of brain cancer in obese/non obese patients. In this paper, we are using 28 attributes (containing patient medical and physical examination results) as an input in our method to get more accurate and precise results related obese and non-obese brain cancer patient detection. One data mining classification technique of Decision trees (C5.0 algorithm were applied). From results it has been seen that C5.0 algorithm provide accurate results. This research showed that data mining techniques can be efficiently
used to predict heart brain cancer tumor in obese/non obese patients. The results which we concluded from this method can be used as supporting or assistant tool for neuro-oncologist for identification and diagnosis of brain tumor. The model which we are using in our research has a high accuracy rate. Due to this high rate of accuracy is making it a supporting tool for helping neuro-oncologist in finding brain tumor in patients. For identifying patients who have a high chances of having the cancer and then transferring those patients to further analysis for confirming the presence of brain tumor and starting its treatment. This system can be further enhanced by adding results and values of more number of input attributes. For prediction and identification of brain cancer we used other data mining techniques e.g. clustering, time series, association rules etc. The text and image mining methods can be used to finding new patterns and important information from huge unstructured medical data.

9. References

1. Frawley WJ, Shapiro GP, Matheus CJ. Knowledge discovery in databases: An overview. Al Magazine. 1996; 13(3):57–70.
2. Rajendran P, Madheswaran M. An improved image mining technique for brain tumor classification using efficient classifier. International Journal of Computer Science and Information Security (IJCSIS). 2009; 6(3):107–16.
3. Naik J, Patel S. Tumor detection and classification using decision tree in brain MRI. International Journal of Engineering Development and Research. 2013; 49–53.
4. Gondal AH, Khan MNA. A review of fully automated techniques for brain tumor detection from MRI. I J Modern Education and Computer Science. 2013; 5(2):55–61. Crossref
5. Ramachandran P, Girija N, Bhuveneswari T. Early detection and prevention of cancer using data mining techniques. International Journal of Computer Applications. 2014; 97(13):48–53. Crossref
6. Padma A, Sukanesh R. Automatic classification and segmentation of brain tumor in CT images using optimal dominant gray level length texture features. International Journal of Advanced Computer Science and Applications. 2011; 2(10):53–9.
7. Ratan R, Sharma S, Sharma SK. Brain tumor detection based on multi parameter MRI image analysis. ICGST-GVIP Journal. 2009; 9(3):9–17.
8. Marcela XR, Agma JMT, Caetano T, Paulo MAM. An association rule-based method to support medical image diagnosis with efficiency. IEEE Transactions on Multimedia. 2008; 10(2):277–85. Crossref
9. Peng H, Long F, Ding C. Feature selection based on mutual information: Criteria of max dependency, max_relevance and min_redundancy. IEEE Transaction on Pattern Analysis and machine Intelligence. 2005; 27(8):1226–38. PMid:16119262. Crossref
10. Edeki C, Pandya S. Comparative study of data mining and statistical learning techniques for prediction of cancer survivability. Mediterranean Journal of Social Sciences. 2012; 3(14):49–56.