# Brain Cancer Risk Prediction Tool Using Data Mining

Tasnuba Jesmin
Department of Information and Communication Technology, Mawlana Bhashani Science and Technology University, Santosh, Tangail-1902

Kawsar Ahmed
Department of Information and Communication Technology, Mawlana Bhashani Science and Technology University, Santosh, Tangail-1902

Md. Zamilur Rahman
Department of Information and Communication Technology, Mawlana Bhashani Science and Technology University, Santosh, Tangail-1902

Md. Badrul Alam Miah
Department of Information and Communication Technology, Mawlana Bhashani Science and Technology University, Santosh, Tangail-1902

## ABSTRACT
Cancer Detection is still challenging for the upgraded and modern medical technology. Evan now the actual reason and total curing procedure of cancer is not invented .After researching a lot statistical analysis which is based on those people whose are affected in brain cancer some general Risk factors and Symptoms have been discovered. The development of technology in science day night tries to develop new methods of treatment. According to a developing country like Bangladesh it is very difficult to bear hug amount of cost for treatment of brain cancer. But it is very easy to protest brain cancer before affected and reduce treatment cost. But the number of brain cancer patients is increasing rapidly in Bangladesh lack of education, money and consciousness. Dreadful, costly and fatal brain cancer also depends on some factors that are known risk factors of brain cancer like other cancers. The detection of Skin Cancer from some important risk factors is a multi-layered problem. Initially according to those risk factors 150 people's data is obtained from different diagnostic centre which contains both cancer and non-cancer patients' information and collected data is pre-processed for duplicate and missing information. After pre-processing data is clustered using K-means clustering algorithm for separating relevant and non-relevant data to Brain Cancer. Next significant frequent patterns are discovered using Pattern Decomposition algorithm shown in Table 1. Finally implement a system using java to predict Brain Cancer risk level which is easier, cost reducible and time saveable.

## General Terms
Computer Science, Data Mining, Bioinformatics.

## Keywords
Brain Cancer, Data Pre-processing, Disease Diagnosis, Classification, K-means clustering, significant frequent pattern and Pattern Decomposition algorithm.

## 1. INTRODUCTION
There is a word that till now in world the brain is the faster processor which sense a problem, send order and commit it within 1 mill second. Physically the brain is soft and spongy collection of tissue. It is covered by [1]:

• The hardy bones of the skull

• Three thin layers of tissue (meninges)

• Cerebrospinal fluid, like Watery fluid that flows through the whole spaces between the meninges and ventricles within the brain.

The brain commits the things what the human wish (like walking and talking) and the things our body does unconsciously (like breathing). The brain is also in charge of our senses (sight, hearing, touch, taste, and smell), memory, emotions, and personality.

Brain has a complex networking System. The medium of network is nerves. As like as wire it carries messages back and forth between the brain and the rest part of the body. Some nerves are directly connected between the brain and eyes, ears, and other parts of the head. Other nerves run through the spinal cord to connect the brain with the rest of parts of human body.

The three major parts of the brain control different activities [1]:

• Cerebrum: The cerebrum collects information from human senses to tell what is going on around us and order to respond according to the circumstances. It controls reading, thinking, learning, speech, and emotions.

• Cerebellum: The cerebellum controls total balance for walking and standing, and other complex actions of human body. That means it controls the physical part of the body.

• Brain stem: It maintains the most important controls. It controls breathing, body temperature, blood pressure, and other basic body functions which are most essential to live.

But our most important part of body (example: brain) without which we can't think any moment or any second to live. Unfortunately sometimes our vital part of body is affected by cancer because of our unconsciousness.

Brain cancer is a disease of the brain in which cancer cells (malignant) arise in the brain tissue. Cancer cells grow to form a mass of cancer tissue (tumor) that interferes with brain functions such as muscle control, sensation, memory, and other normal body functions.

Cancer cells that develop from brain tissue are called primary brain tumors while tumors that spread from other body sites to the brain are termed metastatic brain tumors. Statistics suggest that brain cancer occurs infrequently and is likely to develop in about 22,000 new people per year with about 13,000 deaths as estimated by the National Cancer Institute (NCI) [1].

Treatment for brain cancer depends on the type and stage of the disease, the size and place of the tumor, and your general health and medical history. In most cases, the goal of treatment is to

remove or destroy the cancer completely. Most brain cancers can be cured if found and treated early.

Brain cancer treatment may damage healthy cells and tissues, unwanted side effects sometimes occur. Side effects depend mainly on the type and extent of the treatment. Side effects may not be the same for each person. Before treatment starts, your health care team will tell you about possible side effects and suggest ways to help you manage them. Many brain cancers can be removed quickly and easily. But some people may need supportive care to control pain and other symptoms, to relieve the side effects of treatment, and to help them cope with the feelings that a diagnosis of cancer can bring.

Side effects of brain cancer treatment vary with the treatment plan like surgery, chemotherapy, or radiation. Surgical side effects include an increase in current symptoms, damage to normal brain tissue, brain swelling, and seizures. Other changes in brain functions such as muscle weakness, mental changes, and decreases in any brain-controlled function can occur. Chemotherapy and Radiation therapy usually affect or damage or kill rapidly growing cancer cells but also can affect normal tissue. Common side effects of chemotherapy and Radiation therapy are nausea, vomiting, hair loss, and loss of energy [2].

## 2. BACKGROUND

Cancer is a complex, multiscale process in which genetic mutations occurred. In other word cancer is garbage of dead cells. When the old tissue does not die in natural way then it becomes garbage in body and blocks the space which is for new tissue. This garbage causes the tumor and after one phase it becomes the reason of cancer [1].

Cancer cells that develop in a body organ such as the lung (primary cancer tissue type) can spread via the bloodstream or lymphatic system to other body organs such as the brain. Tumors formed by such cancer cells that spread (metastasize) to other organs are called metastatic tumors. Metastatic brain cancer is a mass of cells (tumor) that originated in another body organ and has spread into the brain tissue. Metastatic tumors in the brain are more common than primary brain tumors. They are usually named after the tissue or organ where the cancer first developed (for example, metastatic lung or breast cancer tumors in the brain, which are the most common types found) [3].

When one got cancer, it's natural to wonder what may have caused the disease that means what is the risk factor. A risk factor is anything that increases a person's chance of developing a brain tumor. Although risk factors often influence the development of a brain tumor, most do not directly cause a brain tumor. Some people with several risk factors never develop a brain tumor, while others with no known risk factors do. However, knowing the risk factors and talking about them with doctor may help anybody make more informed lifestyle and health care choices. The diagnosis of Brain Cancer is a tedious and important task. The detection of Brain Cancer from some important risk factors is a multi-layered problem [2].

Most of the time, the cause of a brain tumor is unknown, but the following factors may raise a person's risk of developing a brain tumor like Age, Gender, Have Taken Any Camo Therapy, Family History, Skin Color, Affected any cancer before, Cell phone use, Smoking, Childhood serious experience, Working period in industries etc. Those are the brain cancer assessment in Bangladeshi population. Those brain cancer risk factors are collected with the help of some people and study.

Brain tumors are more common in children and older adults, although people of any age can develop a brain tumor. Children less 10 years old and people of age 25 to 55 have more opportunity to affect by brain cancer.

In general, men are more likely than women to develop a brain cancer. But some specific types of brain cancers or tumors like meningioma are more common in women.

A person who was affected by any kind of cancer has an increased risk of developing another brain cancer of any type. A person who has two or more close relatives (mother, father, sister, brother, or child) who are responsible for developing brain cancer has a risk factor of developing brain cancer for his own. Rarely, members of a family will have an inherited disorder that makes the brain more sensitive and increases the risk of brain cancer. About 5% of brain tumors may be linked to hereditary (genetic) factors or conditions [4].

Another risk factor of Brain cancer as well as other diseases is taking any camo therapy. A person who has taken any therapy is responsible for occurring different kinds of disease compare to other people who don't take any therapy [2].

In the world as well as Bangladesh White people are more likely to develop gliomas but less likely to develop meningiomas than black people. The people who work in any industry that has radiation, for a long period of time may increase the risk of developing a brain tumor, although there is not yet scientific evidence that supports this possible link.

Now a day the number of cell phone user and smoker is increasing in Bangladesh. Smoking is the vital risk factor of lung cancer as well as brain cancer. Cell phones that have more radiation are more responsible of occurring brain cancer. The World Health Organization (WHO) recommends limiting cell phone use and promotes the use of a hands-free headset for both adults and children.

Now by using those risk factors try to implement a tools using data mining. A widely recognized formal definition of data mining can be defined as "Data mining is the non- trivial extraction of implicit previously unknown and potentially useful information about data" [5]. Data mining has some fields to analysis of data such as classification, clustering, correlations, association rule etc. Now-a-days data mining has been used rapidly by many organizations. In-healthcare, data mining is becoming increasingly popular [6]. Data mining provides the methodology and technology to analysis the useful information of data for decision making.

Data pre-processing is a tedious task of data mining. It mainly used for making analysis appropriate and also making data appropriate for clustering by deleting duplicate records and supplying missing data according to past recorded data. The main benefits of data pre-processing is to reduce memory.

Clustering is a process of separating dataset into subgroups according to the unique feature. Clustering separated the dataset into relevant and non-relevant dataset to brain Cancer. AprioriTid [7] and Decision Tree algorithm [8] are mainly used to find out frequent patterns of dataset. But here use Pattern Decomposition algorithm [9] that is very easy and effective to find out frequent patterns. Frequent patterns, the sets of data are frequently occurred into data warehouse. Significant frequent pattern, the set of data are mostly responsible to brain Cancer. Using this significant pattern we implemented a prediction system for brain cancer that is capable to provide brain cancer risk level.

Day by day the number of brain cancer person is increasing rapidly because of unconsciousness. The Objective of this work is to contract such a tool which can tell people about his/her

approximate condition about brain cancer ,that is he or she in risk or not and how much?

# 3. METHODOLOGY

Collected data is pre-processed to fit the implemented tool appropriately. Here discuss about the whole process (data collection to significant pattern find out) of discovering brain cancer risk level.

## 3.1 Data Collection

Some persons' data is collected from different diagnostic centre where present both male and female information of different age. Obtained data also contains both patient and non patient data. From the previous studies 16 risk factors were considered for brain cancer assessment in Bangladeshi population, which includes- age, gender, hereditary, cell phone use, working period in any industries, color of body/skin, serious experienced in childhood, previous health examination, use of anti-hypersensitive drugs, smoking, food habit, obesity, genetic risk, environment, excessive alcohol and radiation therapy.

## 3.2 Data Pre-processing

Sometime obtained data contains double or more information of same person or missing any values of information. So data pre-processing is an essential which is the vital task of data mining. The main goal of data pre-processing is making an appropriate analysis and suitable for clustering of collected data. Data pre-processing avoids the double data and adds the missing values according to the past recorded data. It also reduces the memory and normalizes the values that are stored in database.

## 3.3 Clustering of Collected Data

The process of categorizing of collected data into different subgroups where each groups have an identical feature is called clustering [10]. Clustering is another vital term of data mining. The clustering problem has been addressed in numerous contents besides being proven beneficial in many applications [11]. The aim of clustering is to classify objects or data into a number of categories or classes where each class contains unique feature. The main benefit of clustering is that the data object is assigned to an unknown class that have unique feature and reduces the memory.

The K-means clustering [12] is a widely recognized clustering tool that is used for robotics, diseases and artificial intelligence application purposes [11]. Here k is a positive integer holding the number of clusters. The pre-processed data is clustered using the K-means clustering algorithm with the value of k=2 which represents two clusters where one cluster contains relevant data to Brain Cancer and another contains remaining data that means non relevant data.

## 3.4 Discover Frequent Pattern

Discover frequent pattern is the most useful, significant and tedious topics of data mining. It is known as the principle data mining problem that intends to find out the frequent items or patterns from the data warehouse [12]. There are different kinds of algorithms, used to mine interesting frequent patterns from databases like association rules, clusters, classifications and correlations etc such as Apriori, AprioriTid, Decision Tree, and FP-Tree.

After clustering pattern decomposition algorithm is used to mine the frequent patterns. The pattern decomposition algorithm is efficient algorithm than AprioriTid and Decision Tree algorithms of extracting the frequent patterns from clustered dataset. The Pattern Decomposition algorithm is the efficient algorithm of extracting the frequent patterns from clustered dataset. Figure-1 shows the execution times for different minimum support. And see that PD is about 30 times faster than AprioriTid with minimal support at 2% and about 10 times faster than AprioriTid at 0.25% [9].

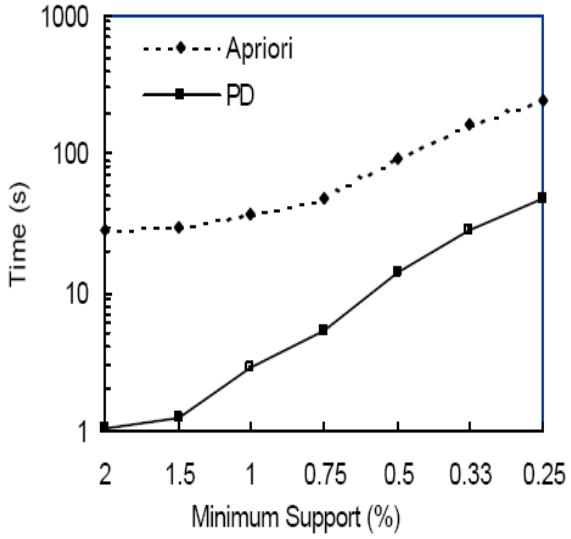


**Fig 1: Comparison with AprioriTid and Pattern Decomposition Algorithm**

**Pseudo code for Pattern Decomposition Algorithm** [9]**:**

PD ( transaction-set $T$ )

1: $D_1 = \{\langle t, 1\rangle | t ? T\}$; $k$=1;
2: while ($D_k$? $F$) do begin
3: forall $p$ ? $D_k$ do // *counting*
4: forall $k$-itemset $s$ ? $p.IS$ do
5: Sup($s D_k$) += $p.Occ$;
6: decide $L_k$ and ~$L_k$;
//*build* $D_{k+1}$
7: $D_{k+1}$ = PD-rebuild($D_k$, $L_k$, ~$L_k$);
8: $k$++;
9: end
10:Answer = ? $L_k$

PD-rebuild ($D_k$, $L_k$, ~$L_k$)

1: $D_{k+1} = F$; $ht$ = an empty hash table;
2: forall $p$ ? $D_k$ do begin
3: // $q_k$, ~$q_k$ can be taken from previous counting
$q_k = \{s | s ?\ p.IS \cap L_k\}$; ~$q_k = \{tt ?\ p.IS \cap$ ~$L_k\}$
4: $u$ = PD-decompose($p.IS$, ~$q_k$);
5: $v = \{s ?\ u |$ $s$ is $k$-item independent in $u\}$
6: add $\langle u\text{-}v, p.Occ\rangle$ to $D_{k+1}$;
7: forall s ? $v$ do
8: if $s$ in $ht$ then $ht.s.Occ$ += $p.Occ$;
9: else put $\langle s, p.Occ\rangle$ to $ht$;
10: end
11: $D_{k+1} = D_{k+1}$ ? $\{p ?\ ht\}$;

The cluster that holds most relevant data to brain cancer is fed as input to pattern decomposition algorithm to mine the frequent patterns present in it. Then the significance weightage of each pattern is calculated using the following approach that will be described in the below subsection.

## 3.5 Significant Pattern Find-Out

After discovering the frequent patterns using pattern decomposition algorithm, the weightage significant patterns are mined by using the Equation (1) [11]

$$\mathbf{S}w(i)=\sum(Wi*Fi) \quad (1)$$

Where $W_i$ represents the weightage of each attribute and $F_i$ holds the number of frequency for each rule.

And significant Frequent Pattern is selected by using the following Equation (2)

$$\mathbf{S}FP=Sw(n) \geq \varphi \text{ for all values of } n \quad (2)$$

Where **S**FP shows significant frequent pattern and φ holds significant weightage.

# 3. RESULTS

Cancer can endanger life and is the leading cause of death in many countries. Brain cancer is the most prevalent of all cancers, and it's increasingly on the rise in Bangladesh. Skin cancer has been a growing problem in Bangladesh because of unconsciousness. The experimental results are separated into two sections. One holds significant frequent patterns discovering part and another prediction tools to brain Cancer.

## 4.1 Result for significant frequent pattern

Using data from data warehouse, the significant patterns are extracted for brain cancer prediction. Avoiding duplicate records and adding missing values collected data are pre-processed and then clustered using K-means cluster algorithm with k=2. Finally significant frequent patterns are mined using Pattern decomposition algorithm shown in Table 1.

**Table 1: Significant Pattern and their corresponding Weightage value using MAFIA Algorithm**

| Significant Patterns | Weight age |
|---|---|
| Age -Sex-Cell Phone Use–Affected family members - Childhood serious experienced - Working period in industries – Affected any cancer before | 145.50 |
| Age- Radiation therapy-Sex- Affected family members - Childhood serious experienced - Working period in industries - Smoke | 130.50 |
| Age- Green foods – Skin color –Sex- Affected family members – Childhood serious experienced – Working period in industries | 125.50 |
| Cell Phone Use – Sex- Sin Color – Affected family members –Green foods- Affected any cancer before | 120.00 |
| Cell Phone Use–Smoke- Childhood serious experienced – Green foods-Affected any cancer before | 115.50 |

## 4.2 Result for Prediction to Skin cancer

Finally brain Cancer prediction tool is implemented using the significant patterns that are collected by using pattern decomposition algorithm. The frequent pattern parameters and their corresponding score are shown at Table 2 and Figure-1 represents the risk level of brain cancer which is implemented using Table 2.

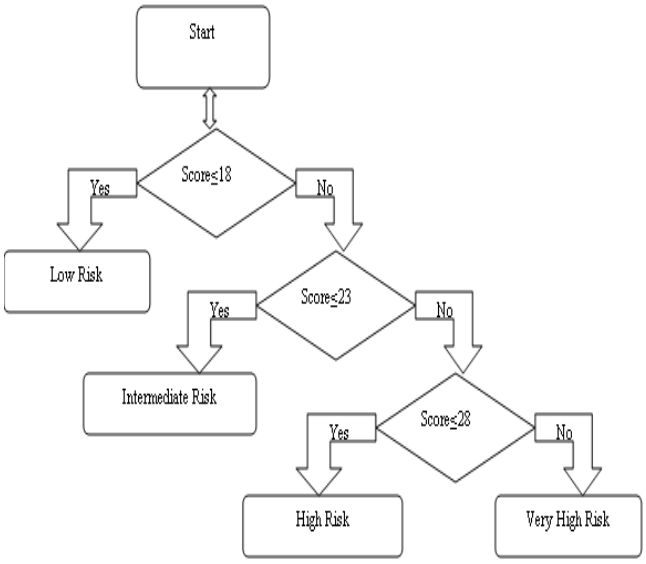


**Fig 1: Flow diagram of decision tree algorithm**

**Table 2: Significant Pattern and their corresponding Weight**
**Age and Score**

| Parameters | Weight age | Score |
|---|---|---|
| Age | Age≤10<br>10<Age≤24<br>24<Age≤45<br>Age>45 | 3<br>1<br>3<br>4 |
| Sex | Male<br>Female | 2<br>1 |
| Cell Phone Use (hours)? | None<br>Less than 5<br>Less than 7<br>More than 7 | 1<br>2<br>3<br>4 |
| Affected Family members? | No<br>Lass than 2<br>More than 2 | 1<br>2<br>3 |
| Working period in industries? | No<br>Less than 5 hours<br>More than 5 hours<br>More than 7 hours | 1<br>2<br>3<br>4 |
| Color of Skin? | Black<br>Swart<br>White | 1<br>2<br>3 |
| Eat green foods/ vegetables? | No<br>Yes | 2<br>1 |
| Smoking? | Yes<br>No | 3<br>1 |
| Childhood serious experienced? | No<br>Yes | 1<br>2 |
| Radiation Therapy in Brain Area? | Yes<br>No | 3<br>1 |
| Affected any cancer before? | Yes<br>No | 3<br>1 |

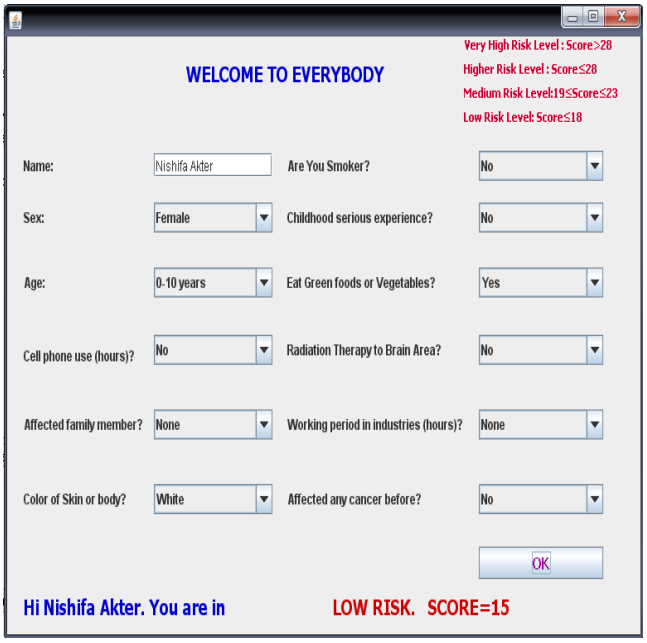


**Fig 2: Brain cancer prediction with Low risk level**

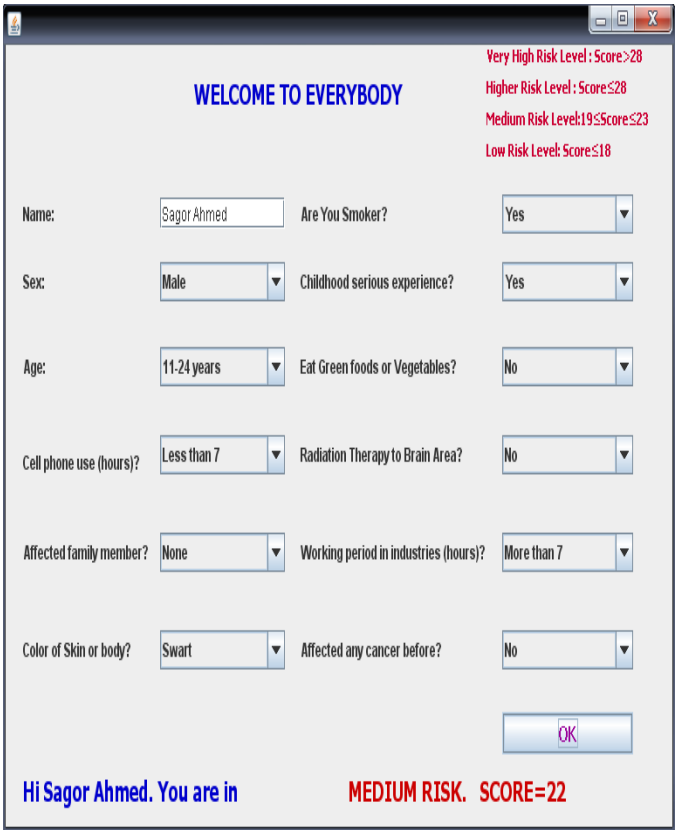


**Fig 3: Brain cancer prediction with Medium risk level.**

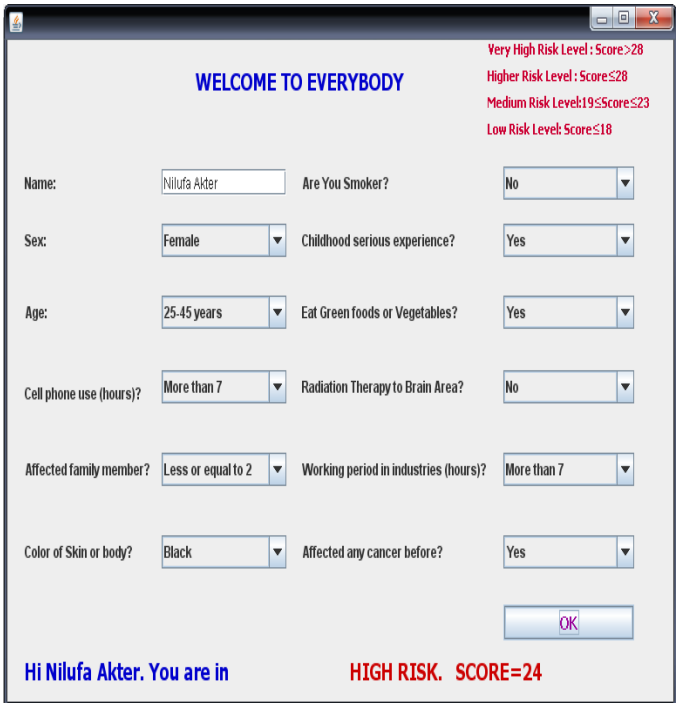


**Fig 4: Brain cancer prediction with High risk level**

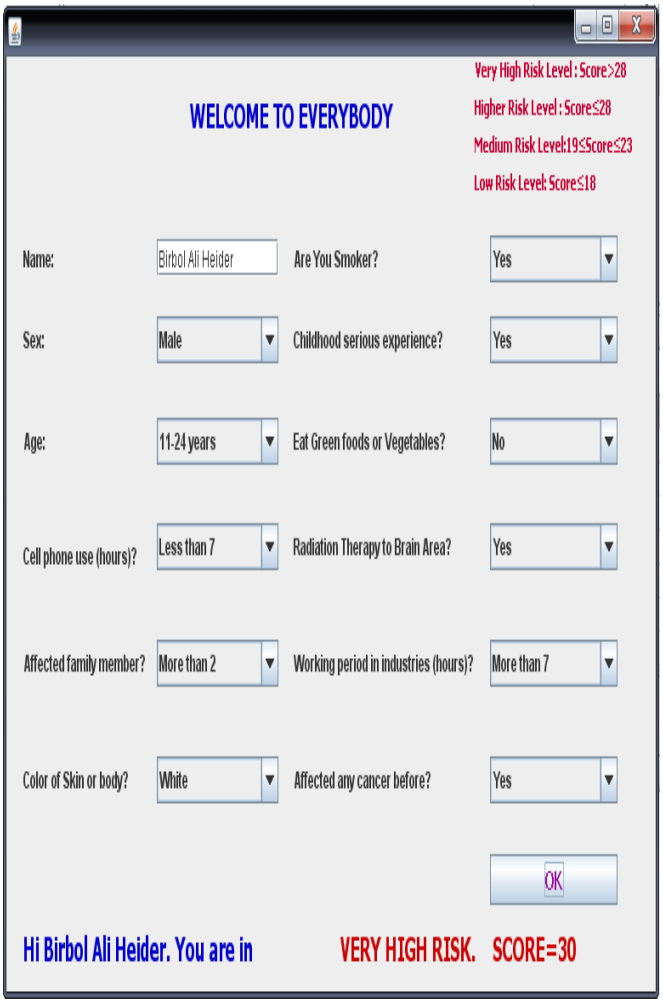


**Fig 5: Brain cancer prediction with Very High risk level**

## 4. CONCLUSION

According to a developing country like Bangladesh it is very difficult to bear hug amount of cost than to solve this deathful disease before affected. Many of them are uneducated, poor and do not even know they have brain cancer. So the ability to predict brain cancer with minimum cost plays an important role in the diagnosis process. The development of technology in science day night tries to develop new methods of treatment. Now-a-days the number of smoker and mobile using person is increasing at alarming rate. With respect to Bangladesh, radiation from mobile is most one risk factor of causing brain cancer. So the possibility of causing Brain Cancer of Bangladeshi as well as whole world is increasing because of unconsciousness. By thinking those are tried to implement a tool that can be able to find out risk level of fatal, deadly, disabling and costly Brain cancer disease. It will be a great achievement to provide any tool that can help to find risk level of Brain Cancer. In this paper proposed an effective brain cancer prediction system based on data mining. The proposed method is implemented using java. The proposed method can efficiently and successfully predict the brain cancer. And will be provided this implemented software through online so that any person can easily check their brain cancer risk level.

## 5. REFERENCES

[1]http://www.medicinenet.com/brain_cancer/page2.htm#what_is_brain_cancer last accessed 12th December 2012.

[2] U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES, National Institutes of Health, NIH Publication No.10-7625, 2010, pp.1-55.

[3] National Library of Australia Cataloguing–in–Publication data: Lifestyle and cancer: knowledge, attitudes and behavior in NSW 2009 SHPN (CI) 120203 ISBN 978-1-74187-760-1, Published by the Cancer Institute NSW, 2012, pp.1-29.

[4] http://www.cancer.net/all-about-cancer last accessed 10-12-12.

[5] Frawley and Piatetsky-Shapiro, Knowledge Discovery in Databases: An Overview. The AAAI/MIT Press, Menlo Park, C.A, 1996.

[6] Hian Chye Koh and Gerald Tan, "Data Mining Applications in Healthcare", Journal of Healthcare Information Management, Vol. 19, No. 2, pp. 64-72.

[7] Dr. Ilias Petrolias and Quan Nguyen**, "**Association rule tool an implementation of AprioriTID Algorithm", ID 2429851.

[8] Jiang Su and Harry Zhang, "A Fast Decision Tree Learning Algorithm", American Association for Artificial Intelligence, 2006

[9] Qinghua Zou, Wesley Chu, David Johnson and Henry Chiu, **"**Pattern Decomposition Algorithm for Data Mining Frequent Patterns", pp. 1-15.

[10] Zakaria Nouir, Berna Sayrac, Benoît Fourestié, Walid Tabbara, and Françoise Brouaye, "Generalization Capabilities Enhancement of a Learning System by Fuzzy Space Clustering," Journal of Communications, Vol. 2, No. 6, pp. 30-37, November 2007.

[11] Shantakumar B.Patil and Y.S.Kumaraswamy, "Intelligent and Effective Heart Attack Prediction System Using Data Mining and Artificial Neural Network", European Journal of Scientific Research ISSN 1450-216X, Vol.31, No.4 , pp.642-656, Inc. 2009.

[12] C. Ordonez, "Programming the K-Means Clustering Algorithm in SQL," Proc. ACM Int'l Conf. Knowledge Discovery and Data Mining, pp. 823-828, 2004.