Journal of Engineering Science 13(1), 2022, 09-20
DOI: https://doi.org/10.3329/jes.v13i1.60558

JES
an international Journal

POTENTIAL RISK FACTORS AND ASSOCIATION OF SIGNIFICANT FACTORS OF BLOOD CANCER IN BANGLADESH USING DATA MINING TECHNIQUES

Farzana Tasnim[1], Tasniya Ahmed[2], Kawsar Ahmed[3], Md. Fazlul Karim Patwary[4] and Md. Karam Newaz[5]

[1,5]Gono Bishwabidyalay, Savar, Dhaka-1344, Bangladesh
[2]Noakhali Science & Technology University, Nokhali-3814, Bangladesh
[3]Mawlana Bhashani Science and Technology University, Tangail-1902, Bangladesh
[4]Jahangirnagar University, Savar, Dhaka-1342, Bangladesh

Received: 09 September 2021 Accepted: 21 April 2022

ABSTRACT

Cancer is now one of the leading causes of death globally and in Bangladesh. This research is focused on the blood cancer patients because there are only a few studies in Bangladesh on blood cancer. Initially, we collected 340 data (blood cancer and non-blood cancer) from BSMMU hospital. Then we use data mining to rank 30 factors associated with blood cancer. Our data mining methods include classification, chi-square (χ^2) test, P-value, and association rule mining. The most potent predictors (P-value < .001) were muscle pull, inability to control the bladder, unusual bleeding, fever/raised temperature, and weakness in the legs. To predict an association between these significant elements, popular rule mining algorithms such as Apriori or Tertius are used. The results of the experiment show that weakness in the legs, fever and inability to control the bladder are common rules of blood cancer. Again, unusual bleeding, rapid illness, and muscle pull are likely to be associated. A fair skin tone with rapid breathing and leg weakness may also be a danger.

Keywords: *Association Rule Mining, Bangladesh, Blood Cancer, Data Mining Techniques, Leukemia, P-Value, Socio-economic.*

1. INTRODUCTION

Cancer is one of the most common causes of death, with blood cancer accounting for 1/3 of all cases. Blood cancers affect the production and feature of our blood cells. The most common types of blood cancers that affect people are lymphoma, myeloma, and leukemia. In maximum blood cancers, the average blood cellular improvement manner is interrupted via the out-of-control growth of a bizarre type of blood cells. These abnormal blood cells prevent our blood from performing many functions, like fighting off infections or preventing severe bleeding. There is no population-based cancer registry in Bangladesh. As a result, the real condition of cancer is generally anonymous here (Hossain et al. 2016). Consistent with World Health Organization (WHO), Bangladesh is facing a growing quantity of the latest cancer cases. The number of the latest subjects is projected to grow by way of approximately seventy-seven percent in 2030. WHO predicts that the wide variety of blood-associated most cancers instances could increase utilizing about forty-eight percent in much less developed countries with the aid of 2030 in comparison to 2012 (Hossain et al. 2014).

Many variables contribute to the occurrence of blood cancer. Although an individual's disease risk is in part determined by his/her genome as well as other non-modifiable factors such as age, ethnicity, and the medical history of the family, there are additional factors that are increasingly being recognized as additional pieces that complete this puzzle (Tan and Naylor, 2022). Cancer is influenced by the same risk variables as other non-communicable diseases: cigarette use, alcohol consumption, physical activity, and diet. These factors are associated with lifestyle choices and environmental influences. These are known as risk factors. These include physical activity, food, weight control, smoking habit, pollution, and infections (Stein & Colditz 2004). Several risk factors have been identified for a variety of hematological cancers (Lee et al., 2015). Alcohol can be considered a cancer risk factor (Rehm et al., 2021). In a hospital population, (Nasir et al., 2015) looked into the incidence and prevalence of blood cancers, and how factors including age, sex, birth order, exact diagnosis, genetic links, and family history influenced the incidence of leukemia. The researchers (Aziz & Qureshi, 2008) identified tiredness, fever, exhaustion, bone pain, pallor, bleeding, and weight loss as prevalent presenting

*Corresponding Author: *rekha.tasnim@gmail.com*

ISSN 2075-4914 (print); ISSN 2706-6835 (online)

symptoms in leukemia. A healthy lifestyle that includes enough physical activity may help to lessen or prevent the long-term effects of pediatric cancer (Rueegg et al., 2012). Environmental risk factors potentially play a significant influence on the growth of pediatric Acute Lymphoblastic Leukemia (ALL) (Pérez-Saldivar, et al., 2016). Researchers (Borugian et al., 2005) used census data to create neighborhood-based income quintiles and stratified the population at risk by sex and 5-year age groups. They found the poorest quintile had a somewhat lower relative risk of childhood leukemia than the richest. (Belson, Kingsley, & Holmes, 2007) concentrated on the demography of leukemia as well as the risk factors linked with the development of ALL or Acute Myeloid Leukemia (AML) in children. Ionizing radiation, nonionizing radiation, hydrocarbons, pesticides, alcohol usage, cigarette smoking, and illicit drug use are among the environmental risk factors highlighted. Clarke et al. (2011) investigated the prevalence of lymphoid malignancies in populations stratified by ethnicity, birthplace, residential neighborhood socioeconomic class, and ethnic enclave status. Consistent with these findings, a hospital population of blood cancer patients was studied to determine whether or not there was a correlation between blood cancer incidence and various factors such as genetic predisposition and socioeconomic status (SES).

In this research, data mining and statistical analysis are taken place. The data mining techniques are popular to search for hidden relationships and frequent item-sets for the prediction of various diseases in the medical field (Shukla et al., 2016). A limited number of publications exist that address blood cancer using the data mining technique. Some of these are the following:

This research (Tayfor et al., 2021) used data mining to classify cancer data into blood cancer and non-blood cancer using pre-defined and post-defined information from blood tests and CT scans. The data mining tool WEKA was used with 10-fold cross-validation to analyze and compare multiple classification algorithms, extract relevant information from the dataset, and accurately find the most suited and predictive model. This study found that Multilayer perceptron has the best ability to predict malignant datasets with an accuracy of 99.3967 %.

(El-Halees et al., 2017) applied data processing techniques to get the relations between biopsy characteristics and blood tumors to predict the disease in an early stage. Using association rules, researchers were able to identify the link between blood test results and blood cancers. Additionally, it displayed the easiest capacity to forecast tumor types of blood disorders with a 79.45% accuracy rate. (Acharya & Kumar, 2019) proposed a work that ambition to survey distinct computer-aided strategies to phase the blood smear image. They have used six hundred images in the experimentation. The model can make a distinction between a regular tangential blood coat and an irregular blood coat. Some papers (Daqqa et al., 2017) used data mining approaches to predict the presence of leukemia in patients by establishing the correlations between blood characteristics and leukemia concerning gender, age, and health condition of patients. They discovered that the DT classifier retrieves properties about outside attributes like a city (eastern regions) that are particularly sensitive to leukemia. There are many cases where data processing techniques are being applied for the diagnosis of various diseases like heart conditions, diabetes, cancer, etc. (Gultepe et al., 2019; Kaur et al.. 2019; Alkaragole et al., 2019).

Blood cancer is rather common in Bangladesh, but nothing is being done about it because of a lack of education and awareness about the problem. Many people of Bangladesh do not even know that they are affected by blood cancer, and in most cases, patients get treatment at the final stage when healing is not feasible. Therefore, early prediction of blood cancer plays a crucial position in the diagnosis technique and prevention approach. Studying the risk factors that lead to blood cancer could help establish policies and therapies to protect against the disease.

The majority of prior research did not describe risk features or the relationship between risk variables and blood cancer. From that context, the purpose of this research is to find out which attributes are an extra concern for predicting blood cancer by evaluating the *P*-value. Also, discover association rules among attributes that assist in causing blood cancer. In this research, we have used the data processing tool SPSS and WEKA 3.8 (Bouckaert et al., 2013). Physicians and patients can use this developed system to quickly know a person's cancer status before screening them for testing cancer.

The rest of the paper is organized as follows: Section 2 describes the detail of the materials and methods for conducting this research. The next section covers experimental results and discussion. Section 4 presents conclusions and future scope. The references are presented at the end of this paper.

2. MATERIALS AND METHODS

This is an ongoing, multi-center, population-based investigation of the relationship between blood cancer-causing variables. The workflow of the research is given in Figure 1. Extensive related works, case studies, and discussions with medical experts and hematologists show that there are several factors influencing blood cancer. These factors are recognized and taken as attributes for this study.

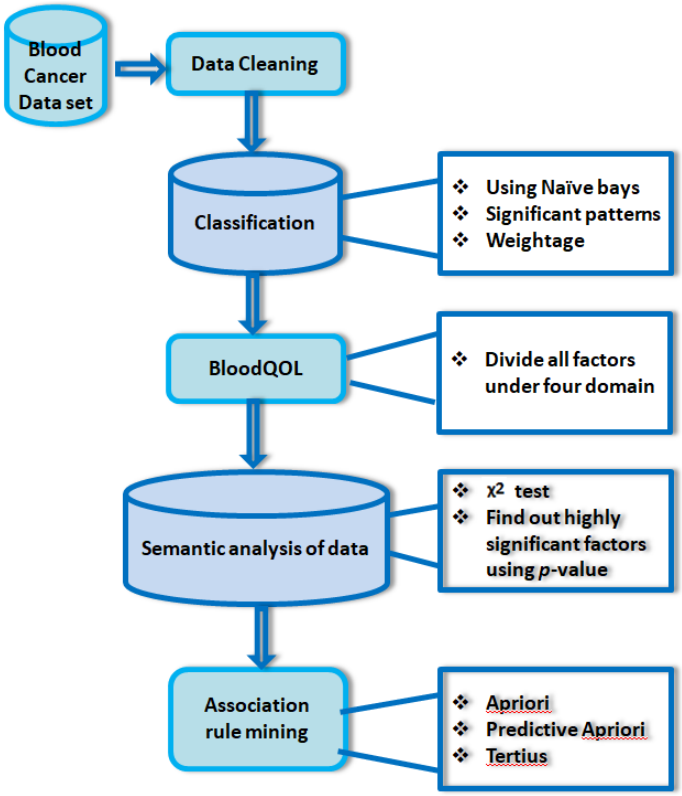

Figure 1: Workflow of this research.

2.1 Data Source

The data for this study was collected from the patient admitted to the National Cancer Institute Bangladesh, Bangabandhu Sheikh Mujib Medical University (BSMMU), and other sources. In recorded data, the participants were both blood cancer and non-blood cancer patient. A total of 170 blood cancer patients and 170 patients with non-blood cancer have been studied for two years. The data consist of more than 30 attributes occupied from previous studies.

2.2 Dataset Description

Attributes of blood cancer used in this research were summarized following different research works (Stein & Colditz, 2004; Lee et al., 2015; Ahmed et al., 2016). The questionnaire sample is provided in a supplementary file. A summarized description of the attributes of the blood cancer data set is given in Table 1.

Table 1: Detailed description of Blood Cancer data set.

Sl. No.	Attribute	Type	Values of attribute	Sl. No.	Attribute	Type	Values of attribute
1	Age	Numeric	10-70 year	16	Fever/raised temperature	Nominal	Yes, No
2	Gender	Nominal	Male, Female	17	Unexpected weight loss	Nominal	Yes, No
3	Living place	Nominal	Town, Village	18	Unusual Bleeding	Nominal	Yes, No
4	Weight	Numeric	40-70kg	19	Fast breathing	Nominal	Yes, No

Sl. No.	Attribute	Type	Values of attribute	Sl. No.	Attribute	Type	Values of attribute
5	Height	Numeric	152-177 cm	20	Stressed behavior	Nominal	Yes, No
6	Being Overweight	Nominal	Yes, No Based on BMI	21	Rapidly becoming more ill	Nominal	Yes, No
7	Skin color	Nominal	Fair, Medium, Dark	22	Night Sweats	Nominal	Yes, No
8	Occupation	Nominal	Student, Business Teacher, Housewife Shopkeeper Job holder, Farmer	23	Not being able to control the bladder	Nominal	Yes, No
9	Job Environment	Nominal	Sunny, Shadow, Both	24	Muscle pull	Nominal	Yes, No
10	Family History	Nominal	Yes, No	25	Mouth sores/ gum edema	Nominal	Yes, No
11	Smoking	Nominal	Yes, No	26	Type 2 diabetes	Nominal	Yes, No
12	Bone/Joint pain	Nominal	Yes, No	27	Eating vegetable	Numeric	No meal per week
13	Pain when moving/coug h	Nominal	Yes, No	28	Previous cancer Treatment	Nominal	Yes, No
14	Difficulty in passing urine	Nominal	Yes, No	29	Blood cancer	Class attribute	Yes, No
15	Weakness usually in legs	Nominal	Yes, No				

2.3 Data Cleaning

Sometimes, collected data from different places contain double or more information about the same person or missing any input values. That may produce a wrong investigation result. So data cleaning is a fundamental step to make a proper analysis of collected records. It decreases memory and normalizes the values that signify information in the form. In our study, we discard 2.9% missing values and 2.98% duplicate values.

2.4 Classification

Naive Bayes is the most effective model to predict patients with cancer disease (Stein & Colditz, 2004). A Naive Bayes classifier is a widely used framework for classification based on a simple theorem of probability known as the Bayes theorem. The Naive Bayes algorithm assesses the chances of the frequencies and combinations of values of a given data set. The probability with vector $x = x1, x2, xn$ is

$$P(h1|xi) = P(xi|h1).P(h1) \,/\, \{P(xi|h1).P(h1) + P(xi|h2).P(h2) \quad (1)$$

Here, $P(h1|xi)$is a subsequent probability, while $P(h1)$ is the prior probability linked with the hypothesis $h1$.

$$P\,(xi) \;=\; \sum_{k=0}^{n} P\,(xi|hi)\;P\,(hi) \quad (2)$$

Thus,

$$P(h1|xi) = P(xi|h1).\,P(h1) \,/\; P\,(xi) \quad (3)$$

The following form is used to discover the major regular pattern.

$$s_{\omega(i)} = \sum_{i=0}^{n} m * t * n \quad (4)$$

$s_{\omega(i)}$ is a Significant Frequent Pattern, m be the number of data, t be the number of attributes, and n be the number of times occurrences of value. We have used Naive Bayes classification because it is easy to use, and it requires a little quantity of training data to approximate the parameters (means and variances of the variables) needed for categorization. The classification result shows that the instances are classified correctly, and there is no incorrectly classified instance. Now we can go for further processing.

2.5 BloodQOL

According to the recent statistical analysis trend named Quality of Life, we categorize data into many classes where every category has exclusive features. Through the Blood specific Quality of Life (BloodQOL) technique (which consists of 31 factors), the data set obtained from the classification is categorized into four domains: personal history domain, Habit domain, family history domain, and disease domain (Ahmed et al., 2016). The factors under these domains are mentioned in Table 2.

Table 2: Socio-demographic quality of life.

Domain	Attribute
Personal History	Name, Age, Gender, height, weight, Being over-weight, Living place, skin color, occupation, job environment
Family History	Family History
Habits	Smoking, Eating vegetable
Disease	Bone/Joint pain, pain when moving or cough, Difficulty in passing urine, Previous cancer treatment, Night sweats, weakness usually in legs, Unexplained weight loss, Fever or raised temperature, Unusual bleeding, Fast breathing, Stressed behavior, Rapidly becoming more ill, Night sweats, Not being able to control the bladder, Muscle pull, Mouth sores, Type 2 diabetes.

2.6 Significant Patterns Find Out

The significance of each attribute under these domains shown in Table 2 is calculated using equation (5) (Ahmed et al., 2016).

*BloodQOL(%)= Response of individual Factor / Total Response of BloodQOL Factors * 100%* (5)

From each domain, we have deleted the attributes whose BloodQOL(%) result was low. The deleted attributes were Name, Height, Weight, Eating vegetables, previous cancer treatment, and Night sweats. The remaining attributes were considered significant.

2.7 Find Out Highly Significant Data

To discover the factors of blood cancer that are highly significant is the objective of this analysis. The P-value of various attributes has been calculated using the χ^2 test formula (Raihan et al., 2017; Nahar et al., 2011; Samarakoon et al., 2019). A Pvalue smaller than .005 is measured as highly significant. This analysis has been implemented by SPSS version 16.0 using the following formula (6).

$$\chi^2 = \Sigma(O - E)^2 \ / \ E \quad (6)$$

Here O = Observed occurrence and E = Expected occurrence

2.8 Association Rule Mining

In recent years, one of the very popular data mining techniques is the association rule, which is applied to discover the frequent pattern among different attributes in medical data. This research uses three trendy algorithms to derive frequent itemsets: Apriori (Nahar et al., 2011), Predictive Apriori (Li et al., 2020), and Tertius (Sarker & Kayes, 2020). For the Apriori algorithm we evaluated rules based on confidence, for the Predictive algorithm we considered accuracy and confirmation for the Tertius algorithm.

3. EXPERIMENTAL RESULTS AND DISCUSSION

3.1 Most Significant Attributes

To calculate the correlation among QOL factors, we use the Chi-square (χ^2) test on the class attribute. Tables 3, 4, 5, and 6 represent the frequency distribution of the attributes. An attribute with *P*-value ≤ .005 can be considered as significant.

The sign * indicates the factors are statistically significant. Table 3 shows that in the personal domain, the job environment factor has the highest impact on BloodQOL as it has the smallest *P*-value (.001), skin color has the

second-highest, and then age, and occupation, respectively. Since the *P*-value of the remaining attributes of Table 3 is more extensive than 0.005, those variables are considered non-significant.

Table 3: P-value of personal domain risk factor attributes.

Attribute		Blood Cancer Status		P-value
		Affected N (%)	Unaffected N (%)	
Age	10 -25	35(21.9)	40(25)	0.003*
	26-40	90(56.2)	35(21.9)	
	41-55	30(18.8)	25(15.6)	
	Above 55	5(3.1)	60(37.5)	
Gender	Male	75(46.9)	110(68.8)	0.064
	Female	85(53.1)	50(31.2)	
Living Place	Town	85(53.1)	115(71.9)	0.098
	Village	75(46.9)	45(28.1)	
Being Over Weight	No	140(87.5)	150(98)	0.057
	Yes	20(12.5)	10(2)	
Skin Color	Fair	110(68.8)	50(31.2)	0.002*
	Medium	40(25)	45(28.1)	
	Dark	10(6.2)	65(40.6)	
Occupation	Student	30(18.8)	20(12.5)	0.005*
	Business	25(15.6)	20(12.5)	
	Teacher	15(9.4)	5(3.1)	
	Shopkeeper	10(6.2)	30(18.8)	
	Housewife	55(34.4)	5(3.1)	
	Job holder	20(12.5)	40(25)	
	Farmer	5(3.1)	40(25)	
Job Environment	Sunny	30(18.8)	60(37.5)	0.001*
	Shadow	75(46.9)	100(62.5)	
	Both Sunny and Shadow	55(34.4)	0(0.0)	

In the family history domain shown in Table 4, the only attribute named family history was found (*P*-value =.005) considerable.

Table 4: P-value of "history domain" risk factor attributes.

Attribute		Cancer Status		P-value
		Affected N (%)	Unaffected N (%)	
Family History	No	150(93.8)	130(87.5)	0.005*
	Yes	10(6.2)	20(12.5)	

As for the disease domain that is indicated in table 5, the most significant features were Bone/Joint pain (*P*-value .003), Weakness in the legs (*P*-value < .001), Fever/raised temperature (*P*-value < .001), Unusual Bleeding (*P*-value < .001), Fast breathing (*P*-value =.005) rapidly becoming more ill (*P*-value = .002), not being able to control the bladder (*P*-value<.001), and muscle pull (*P*-value<.001).

Table 5: *P*-value of "disease domain" risk factor attributes.

Factor		Cancer Status		*P*-value
		Affected N (%)	Unaffected N (%)	
Bone/Joint pain	No	110(68.8)	110(68.8)	0.003*
	Yes	10(6.2)	50(31.2)	
Pain when move/cough	No	100(62.5)	95(59.4)	0.5

Factor		Cancer Status		*P*-value
		Affected N (%)	Unaffected N (%)	
	Yes	60(37.5)	65(40.6)	
Difficulty in passing urine	No	135(84.4)	140(87.5)	0.5
	Yes	25(15.6)	20(12.5)	
Weakness usually in legs	No	10(6.2)	80(50)	0.000*
	Yes	150(93.8)	80(50)	
Fever/raised temperature	No	20(12.5)	90(56.2)	0.000*
	Yes	140(87.5)	70(43.8)	
Unexpected weight loss	No	40(25.0)	75(46.9)	0.059
	Yes	120(75.0)	85(53.1)	
Unusual Bleeding	No	90(56.2)	160(100)	0.000*
	Yes	70(43.8)	0(0.0)	
Fast breathing	No	125(78.1)	120(43.8)	0.005*
	Yes	35(21.9)	90(56.2)	
Stressed behavior	No	90(56.2)	95(59.4)	0.5
	Yes	70(43.8)	65(40.6)	
Rapidly becoming more ill	No	75(46.9)	110(84.4)	0.002*
	Yes	85(53.1)	50(15.6)	
Night Sweats	No	75(46.9)	110(68.8)	0.064
	Yes	85(53.1)	50(31.2)	
Not being able to control bladder	No	155(96.9)	90(56.2)	0.000*
	Yes	5(3.1)	70(43.8)	
Muscle pull	No	75(46.9)	155(96.9)	0.000*
	Yes	85(53.1)	5(3.1)	
Mouth sores/gum edema	No	80(50.0)	125(78.1)	0.018
	Yes	80(50.0)	35(21.9)	
Type 2 diabetes	No	150(93.8)	125(78.1)	0.074
	Yes	10(6.2)	35(21.9)	

This analysis result supported those attributes because the *P*-value was less than .005. The other elements from these factors were discarded. From the habit domain, we found no important factor. So the result table was not shown in this paper.

3.2 Association Rules Among Significant Factors

This segment demonstrates the consequence of applying the association rule to the blood cancer data set. At this point, we prepared a dataset consisting of only major attributes and converted numerical attributes into nominal where needed. To find correlation among these noteworthy factors we applied three association rule algorithms i,e. Apriori, Predictive Apriori, and Tertius in WEKA 3.8 and found several rules for each algorithm. The peak rules were elected based on confidence, accuracy, and confirmation respectively. Here, terms are separated in the rules using the symbol '$\cup$' and '$\cap$' that stands for 'or' and 'and' correspondingly.

Table 6: Association rule mining for blood cancer by the Apriori algorithm.

Serial	Rules	Result (Blood Cancer)	Confidence
	Rules for class attribute Yes		
1	Weakness usually in legs = Yes $\cap$ Fever/raised temperature = Yes $\cap$ Not being able to control bladder = No	Yes	1
2	Family history = No $\cap$ Fever/raised temperature =Yes $\cap$ Fast breathing =No $\cap$ Not being able to control bladder=No	Yes	1

Serial	Rules	Result (Blood Cancer)	Confidence
3	Bone/Joint pain =Yes ∩ Weakness usually in legs = Yes ∩ Not being able to control bladder = No	Yes	1
4	Skin color (Fair /Medium /Dark) = Fair ∩ Weakness usually in legs = Yes ∩ Fever/raised temperature = Yes	Yes	1
5	Age = Young ∩ Weakness usually in legs=Yes ∩ Not being able to control bladder = No Rules for class attribute No	Yes	1
1	Bone/Joint pain = No ∩ Fever/raised temperature = No ∩ Unusual bleeding = No ∩ Muscle pull=No	No	1
2	Fever/raised temperature = No ∩ Unusual bleeding = No ∩ Muscle pull = No	No	1
3	Unusual bleeding = No ∩ Fast breathing = Yes ∩ Rapidly becoming more ill=No	No	.95
4	Fast breathing = Yes ∩ Rapidly becoming more ill = No ∩ Muscle pull = No	No	.95
5	Job environment (Sun / Shadow) = shadow ∩ Unusual bleeding = No ∩ Rapidly becoming more ill = No ∩ Muscle pull =No	No	.95

Apriori algorithm has been applied in weka 3.8 for the class attribute (yes and no) with minimum support of .25 and minimum confidence of 90%. The rules with a confidence level of at least 95% were chosen. Here, only the rules containing class attributes on the right-hand side (RHS) were reported. We found 6 yes rules with 100% confidence, and also found 6 no rules, where 3 rules have 100% confidence and the other 3 rules have 95% confidence (see Table 6).

Then, we executed the Predictive Apriori algorithm and considered the association rules that only predict the class level. We reported a total of 12 rules for this algorithm with the accuracy highest of 99.44% and the lowest at 97.02% (see Table 7).

Table 7: Association rule mining for blood cancer by the Predictive Apriori algorithm.

Serial	Rules	Result (Blood Cancer)	Accuracy (%)
	Rules for class attribute Yes		
1	Weakness usually in legs = Yes ∩ Fever/raised temperature = Yes ∩ Not being able to control bladder = No	Yes	99.44%
2	Bone/Joint pain = Yes ∩ Not being able to control bladder = No	Yes	99.42%
3	Age = Young ∩ Weakness usually in legs = Yes ∩ Not being able to control bladder = No	Yes	99.38%
4	Skin color(Fair/Medium/Dark)=Fair ∩ Weakness usually in legs = Yes ∩ Fast breathing=Yes	Yes	98.56%
5	Age = Young ∩ Skin color (Fair / Medium / Dark) = Fair ∩ Weakness usually in legs = Yes	Yes	99.33%
	Rules for class attribute No		
1	Bone/Joint pain = No ∩ Unusual bleeding = No ∩ Rapidly becoming more ill = No	No	99.38%
2	Fever/raised temperature = No ∩ Unusual bleeding = No ∩ Fast breathing = No	No	99.33%
3	Age = Young ∩ Weakness usually in legs = Yes ∩ Not being able to control bladder = No	No	99.30%
4	Family history = No ∩ Weakness usually in legs = Yes ∩ Fever/raised temperature = No ∩ Unusual bleeding = No ∩ Muscle pull = No	No	99.05%

Serial	Rules	Result (Blood Cancer)	Accuracy (%)
5	Skin color (Fair/Medium/Dark) = Fair ∩ Bone/Joint pain = No ∩ Fever/raised temperature = No ∩ Rapidly becoming more ill = Yes	No	98.56%

Finally, we implemented the Tertius association rule mining algorithm. This algorithm generated so many association rules. We kept rules holding class level on LHS or RHS and which are applicable for plenty of attributes based on confirmation. In this case, the highest confirmation was 84.94% (see Table 8).

Table 8: Association rule mining for blood cancer by the Tertius algorithm.

Serial	Rules	Result (Blood Cancer)	confiration
	Rules for class attribute Yes		
1	Weakness usually in legs = Yes ∩ Not being able to control bladder = No ∪ Age = Younger	Yes	83.08%
2	Unusual bleeding = Yes ∪ Rapidly becoming more ill = Yes ∪ Muscle pull = Yes	Yes	75.31%
3	Job environment (Sun /Shadow) = Both ∪ Unusual bleeding = Yes ∪ Rapidly becoming more ill = Yes	Yes	75.08%
4	Occupation = Job ∪ Unusual bleeding = Yes ∪ Muscle pull = Yes		72.66%
5	Skin color (Fair / Medium / Dark) = Fair or Job environment (Sun / Shadow) = Both ∪ Unusual bleeding = Yes	Yes	63.64%
	Rules for class attribute No		
1	Skin color(Fair / Medium / Dark) = Dark ∪ Fever/raised temperature = No ∪ Not being able to control bladder = Yes	No	84.94%
2	Skin color (Fair / Medium / Dark) = Dark ∪ Fever/raised temperature = No ∪ Age = Old	No	81.59%
3	Weakness usually in legs = No ∪ Fever/raised temperature = No ∪ Not being able to control bladder = Yes	No	79.19%
4	Skin color(Fair / Medium / Dark) = Dark ∪ Weakness usually in legs = No ∪ Fever / raised temperature = No	No	72.04%
5	Job environment (Sun / Shadow) = Sun ∪ Weakness usually in legs = No ∪ Fever/raised temperature = No	No	66.06%

3.3 Discussion

Our first experiment was to find significant attributes based on the *P*-value. The result of this study indicates the predictive role of 13 factors amongst 31 input attributes in the model. Here, we find that always fever/raised temperature, muscle pull, not being able to control the bladder, unusual bleeding, and weakness usually in the legs have greater significance as their *P*-value is < .001. Other important influence elements are job environment, skin color, rapidly becoming more ill, bone/joint pain, and age. The *P*-value of these is ≤ .003 Remaining considerable attributes are family history, fast breathing, and, occupation. Top risk factors are represented in a bar chart in Fig 2.

Our next study was to find an association between these elements. By using three association rule mining algorithms we find the rules shown in tables 6,7 and 8. From the tables, we observe that Apriori and Predictive Apriori algorithm produces more accuracy for class = Yes. on the other hand, the Tertius algorithm gives a more accurate result for class = no instead of class = yes. In general, we see almost the same rules in these three algorithms. However, weakness usually in the legs = yes, fever/raised temperature = yes, and not being able to control bladder = no is a common rule for blood cancer = yes. Again, unusual bleeding = yes, rapidly becoming more ill = yes, and muscle pull = yes rule is likely to have a significant association. Besides, having a fair skin tone with fast breathing and weakness in the legs might be a threat. From Tables 6, 7 and 8 it is confirmed that

people in the age group younger (10-25) and young (26-40) are at higher risk for blood cancer. In Table 5, we see family history as an inversely significant attribute. Here, in association rules, we see that having no family history is correlated with both blood cancer yes and blood cancer no class.

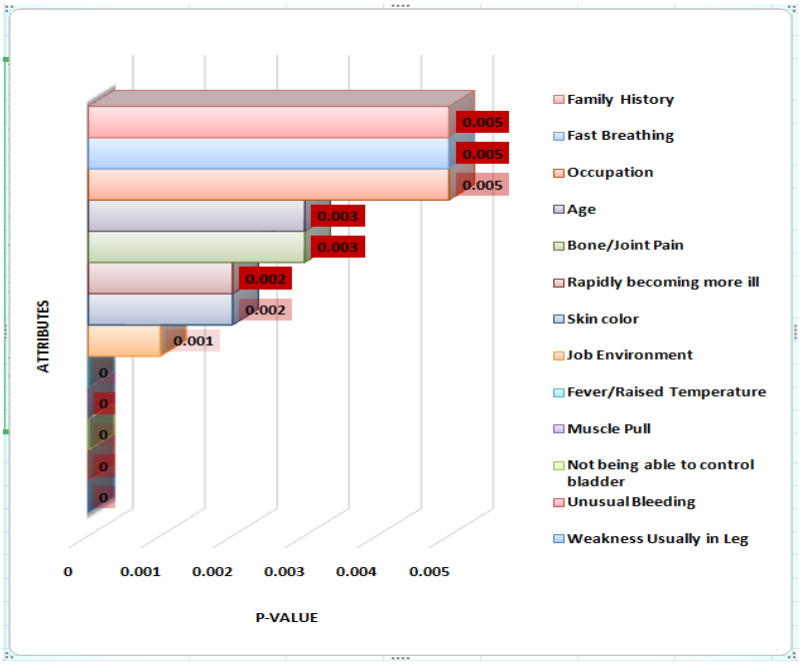

Figure 2: Bar chart of most significant attributes along with the p-value.

While comparing our outcome to older studies, we must point out that those studies find different classification algorithms' accuracy levels. In contrast, our study casts a new light on the significance of attributes. Also relation among those attributes. Some of the limitations of this study are that we conducted a face-to-face interview to collect data. It may include some wrong information by mistake. The frequency of data may change for large volumes of data.

4. CONCLUSION AND FUTURE SCOPE

In recent years, the quantity of blood cancer patients is growing at a rapid pace in Bangladesh. The most efficient way to reduce death by cancer is to detect it earlier. This paper has worked on blood cancer and different data mining tools and techniques which are becoming popular day by day. This paper projected a multi-layered data mining method combining The Naive Bayes classifier, BloodQOL, and association rule to guess blood cancer risk. It will be beneficial for the early detection of blood cancer, which is better than treatment and can diminish cancer passing. This aspect of the research suggested that age, skin color, occupation, job environment, bone/Joint pain, weakness usually in legs, fever or raised temperature, unusual bleeding, fast breathing, rapidly becoming more ill, muscle pull, etc. are strongly linked with blood cancer. This research also suggests an association between these factors. We can use this finding to raise consciousness among natives about these factors that may cause blood cancer. It will be useful for fast prohibition and superior to treatment. Regardless, future research could continue to explore different machine learning feather selection methods to find out significant attributes.

STATEMENTS, DECLARATIONS AND ACKNOWLEDGEMENT

This research proposal won the National Science and Technology (NST) fellowship award from the Ministry of Science and Technology, Bangladesh. The authors acknowledge the Bangabandhu Sheikh Mujib Medical University (BSMMU) hospital authority for permitting the collection of data used in this research. There are no conflicts of interest. "The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results".

REFERENCES

Acharya, V., & Kumar, P. (2019). Detection of acute lymphoblastic leukemia using image segmentation and

data mining algorithms. *Medical & biological engineering & computing*, *57*(8), 1783-1811.
Ahmed, K., Jahan, P., Nadia, I., & Ahmed, F. (2016). Assessment of menopausal symptoms among early and late menopausal midlife Bangladeshi women and their impact on the quality of life. *Journal of menopausal medicine*, *22*(1), 39-46.
Alkaragole, M. L. Z., & Kurnaz, S. (2019). Comparison of data mining techniques for predicting diabetes or prediabetes by risk factors. *International Journal of Computer Science and Mobile Computing*, *8*, 61-71.
Aziz, F., & Qureshi, I. Z. (2008). Clinical and cytogenetic analyses in Pakistani leukemia patients. *Pakistan Journal of Zoology*, *40*(3).Belson, Martin, Beverely Kingsley, and Andrianne Holmes. 2007. "Risk Factors for Acute Leukemia in Children: A Review." *Environmental Health Perspectives* 115(1): 138–45.
Borugian, M. J., Spinelli, J. J., Mezei, G., Wilkins, R., Abanto, Z., & McBride, M. L. (2005). Childhood leukemia and socioeconomic status in Canada. *Epidemiology*, *16*(4), 526-531.
Clarke, C. A., Glaser, S. L., Gomez, S. L., Wang, S. S., Keegan, T. H., Yang, J., & Chang, E. T. (2011). Lymphoid malignancies in US Asians: incidence rate differences by birthplace and acculturation. *Cancer Epidemiology and Prevention Biomarkers*, *20*(6), 1064-1077.
Daqqa, K. A. A., Maghari, A. Y., & Al Sarraj, W. F. (2017, May). Prediction and diagnosis of leukemia using classification algorithms. In *2017 8th international conference on information technology (ICIT)* (pp. 638-643). IEEE.
El-Halees, A. M., & Shurrab, A. H. (2017). Blood tumor prediction using data mining techniques. *Health Informatics—An International Journal*, *6*.
Gultepe, Y., & Rashed, S. (2019). The use of data mining techniques in heart disease prediction. *vol*, *8*, 136-141.
Hossain, M. S., Iqbal, M. S., Khan, M. A., Rabbani, M. G., Khatun, H., Munira, S., ... & Sultana, T. A. (2014). Diagnosed hematological malignancies in Bangladesh-a retrospective analysis of over 5000 cases from 10 specialized hospitals. *BMC Cancer*, *14*(1), 1-7.
Hossain, M. S., Begum, M., Mian, M. M., Ferdous, S., Kabir, S., Sarker, H. K., ... & Karim-Kos, H. E. (2016). Epidemiology of childhood and adolescent cancer in Bangladesh, 2001–2014. *BMC Cancer*, *16*(1), 1-8.
Kaur, P., Pruthi, Y., Bhatia, V., & Singh, J. (2019). Empirical analysis of cervical and breast cancer prediction systems using classification. *International Journal of Education and Management Engineering*, *9*(3), 1.
Lee, K., S. G. Kim, and D. Kim. 2015. "Potential Risk Factors for Haematological Cancers in Semiconductor Workers." *Occupational Medicine* 65(7): 585–89.
Li, D., Yang, D., Zhang, J., & Zhang, X. (2020). Ar-ann: Incorporating association rule mining in artificial neural network for thyroid disease knowledge discovery and diagnosis. *IAENG International Journal of Computer Science*, *47*(1), 25-36.
Pérez-Saldivar, M. L., Rangel-López, A., Fajardo-Gutiérrez, A., & Mejía-Aranguré, J. M. (2016). Environmental Factors and Exposure Time Windows Related to the Etiology of Acute Lymphoblastic Leukemia in Children. In *Etiology of Acute Leukemias in Children* (pp. 207-290). Springer, Cham.
Nahar, J., Tickle, K. S., Ali, A. B. M., & Chen, Y. P. P. (2011). Significant cancer prevention factor extraction: an association rule discovery approach. *Journal of medical systems*, *35*(3), 353-367.
Nasir, M., Jabeen, F., Hussain, S. M., Shaheen, T., Samiullah, K., & Chaudhry, A. S. (2015). Impact of Consanguinity, Environment, Socio-Economic and Other Risk Factors on Epidemiology of Leukemia. *Pakistan Journal of Zoology*, *47*(4).
Raihan, M., Mondal, S., More, A., Sagor, M. O. F., Sikder, G., Majumder, M. A., ... & Ghosh, K. (2016, December). Smartphone-based ischemic heart disease (heart attack) risk prediction using clinical data and data mining approaches, a prototype design. In *2016 19th International Conference on Computer and Information Technology (ICCIT)* (pp. 299-303). IEEE.
Rueegg, C. S., Von Der Weid, N. X., Rebholz, C. E., Michel, G., Zwahlen, M., Grotzer, M., ... & Swiss Paediatric Oncology Group (SPOG). (2012). Daily physical activities and sports in adult survivors of childhood cancer and healthy controls: a population-based questionnaire survey. *PloS one*, *7*(4), e34930.
Samarakoon, Y. M., Gunawardena, N. S., Pathirana, A., Perera, M. N., & Hewage, S. A. (2019). Prediction of colorectal cancer risk among adults in a lower-middle-income country. *Journal of Gastrointestinal Oncology*, *10*(3), 445.
Sarker, I. H., & Kayes, A. S. M. (2020). ABC-RuleMiner: User behavioral rule-based machine learning method for context-aware intelligent services. *Journal of Network and Computer Applications*, *168*, 102762.
Shukla, S., Gupta, D. L., & Prasad, B. R. (2016). Comparative study of recent trends on cancer disease prediction using data mining techniques. *International Journal of Database Theory and Application*, *9*(9), 107-118.
Stein, C. J., and G. A. Colditz. 2004. "Modifiable Risk Factors for Cancer." *British Journal of Cancer* 90(2): 299–303.
Tan, Keely, and Matthew J. Naylor. 2022. "The Influence of Modifiable Factors on Breast and Prostate Cancer Risk and Disease Progression." *Frontiers in Physiology* 13(March).
Tayfor, N. B., & Mohammed, S. J. (2021). A Comparison Study of Data Mining Algorithms for blood Cancer

Prediction. *Passer Journal of Basic and Applied Sciences*, *3*(2), 174-179.