Cancer Disease Prediction Using Naive Bayes And Other Machine Learning Algorithm.

Devanshi Vats, B.Tech CSE,
5th Semester

***Abstract*—Cancer is a rising deadly disease which causes the death of 10% among all the diseases . There are over 100 Types of cancer. Predicting Cancer plays a vital role for progressing data mining applications. Naive Bayes, k-nearest neighbor and j48 algorithm are used in this paper for predicting cancer disease. Naive Bayes is easy to build and really useful for very big dataset. K-nearest neighbor is uses dataset and create a dataset by separated into different classes and also predicting classification of new points. J48 Classifier are based on the decision Tree from training datasets, using the fact that each of them and data sets can be used for decision-making it into smallest subset. Weka tool is used for the purpose of measuring the accuracy of the cancer disease dataset including 09 types of cancer. 10-fold cross-validation is used for predicting cancer disease. In Naive Bayes the accuracy is 98.2%, k-nearest neighbor accuracy is 98.8% and j48 accuracy is 98.5%.**

Keyword— brain cancer; blood cancer; prostate cancer; pancreatic cancer; ovarian cancer; breast cancer; esophageal cancer; lung cancer; colorectal cancer; naive Bayes; k-nearest neighbor; j48; classification; accuracy

I. INTRODUCTION

A. Background

Cancer can grow to 15 million by 2020. Introducing appropriate tools for identifying health practitioners Cancer cases will be an important step in the treatment diagnosis of cancer patients. Cancer is the name given to a collection of related diseases. In all types of cancer, some of the body's cells begin to divide without stopping and spread into surrounding tissues. Cancer may start from anywhere in the body, trillions of cells is made by it. Cancer is a genetic disease—that is, it is caused by changes to genes that control the way our cells cause, especially how they increase and divide. A cancer that has spread from the place where it first started to another place in the body is called metastatic cancer. The process by which cancer cells spread to other parts of the body is called metastasis [5].

B. Motivation

The technique of data mining is increasing very rapidly in the medical field due to its success in the classification and prediction algorithms that helps doctors in Decision making. We are looking for ways to improve patient health and reduce costs of medicine, data mining helps a lot to fulfil this purpose. Many people want to know about the earlier symptoms of cancer so that they can take decision in initial stage [2]. That is the reason behind the research that help people to get idea about cancer and make them conscious. Here we use the Weka abbreviation for Waikato an analysis of knowledge, this is open source data mining program, was newly developed at Waikato University New Zealand, and it is licensed under the GNU General Public License [11]. It helps us to predict cancer disease properly and help us to take a proper decision. Data Mining Prediction Tool for Predict cancer. We used three well known classification algorithm named Naive Bayes, K-Nearest and j48 algorithm. For these purposes we use a dataset which will show us the correct answer. The dataset is verified by the doctors and it contains the data of undergoing cancers patients. 10-fold cross validation is used.

C. Paper organization

This paper contains eight section .Section I contains Introduction .Section II will provide related work, section III will provide problem statements, section IV will provide preliminaries, section V will provide methodology, section VI will provide the result and discussion and section VII will provide conclusion. The references have been attached in the last segment.

II Related Work

Scientists are endeavoring to diminish the passing aftereffects of malignancy, so there are numerous inquiries about foreseeing survivability of disease.

Moataz M. Abdelwahab, Shimaa A. Abdelrahman [1] predicted that Lung cancer is among the most common genetic diseases that thousands of people die from each year. It develops in human body due to damage of normal genes caused by many reasons such as smoking cigarette, which chemically activate the oncogenes and deactivate the tumor suppressor the normal lung cell and produce mutations that result in tumors. The proposed algorithm reflects the genetic information into eigenspace by utilizing 2DPCA technique.

Dona Sara Jacob et al. [8] varied classification algorithms and the clustering algorithm are used. The outcome indicates that the classification algorithms are superior predictors than the clustering algorithms. Studies filtered all algorithms based on D. Support Vector Machine lowest computing time and accuracy and it came up with the conclusion that Naïve Bayes is a superior algorithm compared SVM model is a machine learning technique which is based to decision tree and k-nearest neighbor, because it takes lowest on the statistical learning theory. Most promising clustering algorithm with the accuracy of 68%. F. Expectation Maximization The research shows that the classification algorithms are better predictor than clustering algorithms. Conclusion from the above comparisons, it is concluded that the classification algorithms works better than the clustering algorithms in predicting breast cancer. The best algorithm for predicting breast cancer is purely based on the accuracy of the algorithm.

G.N.Satapathi et al. [10] predicts that Cancer is caused by abnormalities in the genetic material of the transformed cell. Cancer-promoting genetic abnormalities may randomly occur through errors in DNA replication or are inherited and thus present in all cells from birth. The heritability of cancer is usually affected by complex interactions between carcinogens and the host's genome. There is a diverse classification scheme for multiple genomic changes which may contribute to the origination of cancer cells. Most of these changes can be classified as mutations or changes in the nucleotide sequence of genomic DNA. Most cancers originate from random mutations that mature in body cells during one's lifetime either as an error with the inception of cell division or in response to specific injuries from environmental agents such as exposure to radiation or chemicals. Nanotechnology is also adopted to develop accurate and sensitive biomedical devices for cancer genome study.

Ms. Rashmi G D et al. [9] proposed that data mining is the process of retrieving the information from huge data The paper provides analysis of Classification and Prediction data mining techniques. Breast cancer is the regularly found cancer in women. It is found in both men and women. Breast cancer represents 12% of new cancer cases. It is the second most common cancer found worldwide. The breast cancer if detected in early stage is very helpful to identify the type of tumor.

III. Problem Statement

NSCLC is the profoundly announced malignant growth which can be restored by early location and medications. In spite of the fact that there are many progressed computational advancements to help early location of lung diseases none are utilized on account of low unwavering quality and high establishment cost. miRNA is even utilized for rebuilding genomes. In an examination done by Tingting Wang, it is discovered that miRNA can be efficiently modified utilizing dangerous pleural radiation. Countless works have done over miRNA in absolutely necessary, optional and tertiary structure. System of miRNA target communication is followed by Yonghua Wang et al. Utilizing different sub-atomic elements reenactments and thermodynamic examination. In this work, they likewise created non-coding RNA (ncRNA) restricting model. Target Scan is an online database which is utilized to distinguish hereditary focuses of miRNAs. In this database, all the miRNAs and their particular target qualities are recorded [3].

Cigarette smoking has a huge job in changing miRNA articulation. There are numerous miRNAs which have critical job in the force of aviation route barricade in interminable obstructive aspiratory malady (COPD). This should be possible by low pass separating the flag, gave our 'applicable' flag data is contained inside this band. More often than not, this 'band' is of low go in nature, for example speech, natural pictures and so forth [10].

Two-dimensional key part investigation (2DPCA) and two-dimensional direct discriminant examination (2DLDA) are new procedures for face acknowledgment. The primary thoughts behind 2DPCA and 2DLDA are that they be controlled 2D lattices instead of the usual PCA and LDA, which depend on 1D vector. In some writing, there has been a propensity to incline toward 2DLDA over 2DPCA in light of the fact that, as instinct would recommend, the previous

arrangements specifically with separation between classes, though the last arrangements with the information in its completely for the key segments examination without giving a specific consideration to the fundamental class structure. In this paper, to look at the exhibitions of the two techniques, a progression of trials performs on two face picture databases: ORL and CAS-Ring. The trials results demonstrate that the execution of 2DLDA isn't in every case superior to that of 2DPCA. Especially, on account of extensive subjects, 2DPCA can outflank 2DLDA [1].

The long clinical information on malignant growth patients was gathered from various sources. Completed a ton of research work from the web, went to a various medical clinic and examined with a couple of the number of authorities. There are numerous qualities for the programmed conclusion framework.

IV. PRELIMINARIES

A. Data Sources:

In the case of predicting cancer, undergoing patient's data is used. The cancer patients' data collected from different sources. We work a lot to collect data research a lot for creating a dataset. Doctors and experts help us for making the dataset. There are 1059 data and it contains 61 attributes and 1 class attributes. According to 61 attributes, the 1 class attributes is created.61 attribute includes symptoms and some tests part of cancers and the 1 class which represents the types of cancers. The cancer disease dataset format is .csv format. The cancer disease dataset Table I represents the Symptoms part and Table II represents the Tests parts. The description of all those parts symptoms, a test and also each attributes details is in section V.

TABLE I. Symptoms part

Serial Number	Attributes Name
01	Age
02	Gender
03	Vomiting
04	Vision Problem
05	Confusion
06	Seizures
07	Abdominal Pain
08	Dark Urine
09	Jaundice
10	Diabetes
11	Back Pain
12	Frequently urine
13	Decrease Urine
14	Blood Urine
15	Rectum Pain
16	Bone Pain
17	Abdominal Bloating
18	Difficulty Eating
19	Pelvic Pain
20	Fatigue
21	Means Irregular
22	painful Intercourse
23	Lymph
24	Armpits Pain
25	Redness
26	Rash
27	Discharge Blood
28	Inverted Nipple
29	Change Size
30	Flaking
31	Heart Burn
32	Swallowing pain
33	Chest pain
34	Cough
35	Coughing up Blood
36	Shortness of Breath
37	Rectal Bleeding
38	Blood in Stool
39	Diarrhea
40	Stool Change
41	Frequent Infection
42	Night sweats
43	Bleeding

Table I represents the symptoms part. This part contains some symptoms of having cancer. Some common symptoms are found for several types of cancer. These symptoms are used in a dataset and help to take prediction.

TABLE II. Test Part

Serial Number	Test Name
01	Bone Marrow Aspiration
02	Mammogram
03	CT scan
04	MRI
05	Ultrasound
06	Digital Rectal Exam
07	Prostate Specific Antigen
08	Pelvic Exam
09	Pelvic MRI
10	Endoscopy
11	Barium Swallow
12	Sputum Cytology
13	Fecal Occult Blood test
14	Stool DNA test
15	Flexible Sigmoidoscopy
16	Colonoscopy
17	Biopsy

Table II represents the Test part. This part has some test of having cancer. These tests are used in a dataset and help to takes prediction. Several tests are performed but one test is common for all cancer is called biopsy. A biopsy is the only medium to verify cancer disease.

B. Classifiers

The term classification is the most common term in Weka classification is done with the leveled dataset it helps us in decision making purpose. There are two types of classification binary and multi-class targets Binary contains two types of results and another side multi-class target provides greater than two values [9]. Its main purpose is to classify and predict perfectly [9] [12]. Prediction is something based on some related data. It can tell us what will happen in the ahead time. A disease is predicted by using it in data mining. Prediction creates a relationship between a data people know and data that people need to predict for ahead time reference [13]. There are many types of classifiers algorithm. In this paper, three types of classifiers is used Naive Bayes, K-Nearest Neighbor, and J48 algorithm. We worked on windows 10 operating system and Weka 3.6 version.

Naive bayes is a simple technique known as "probabilistic classifiers". It is the combination of algorithms which share common terms where every attribute being classified independently from other featured values [14]. It is carried out in the assumption that the impact of an attribute value on a class does not depend on other attribute values [15]. It is easy to build and really useful for a very big dataset. The posterior probability, P (c|x), from P(c), P(x), and P (x|c).

Naive Bayes assume the effect of a predictor (x) on a given class(c).

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)} \quad \text{.....(i)}$$

$$(c|X) = P(x_1|c) \times P(x_2|c) \times ... \times P(x_n|c) \times P(c) \text{ .... (ii)}$$

Here $(c|X)$ represents Posterior Probability. (x) Indicates Predictor Prior Probability and last $P(c)$ indicates Class Prior Probability.

K-Nearest Neighbor (KNN) is a lazy and non-parametric algorithm. It uses dataset and creates a dataset by separated into different classes and also predicting the classification of new points [11]. K-nearest neighbors also called instance-based learning, which memorizes the observations to classify the unseen test data. It compares the test observations with the nearest training observations.

J48 uses an open source Java implementation C4.5 algorithm [16]. Suppose we have a dataset that contains independent and another contains dependent variables. After applying the decision tree like the J48 algorithm on that dataset allow us to predict new dataset record. It works well on continuous and discrete also missing data values [6]. It also gives an option for pruning trees after creation. The classifier uses the greedy method and reduced the error and create declarations for classification [6].

V. METHODOLOGY

A. Data Statement

Cancer is not a single disease is a disease of a group which includes abnormal growth of cells. Here 9 types of cancer works is represented which is well known. Generally analysis the symptoms and then also analysis some tests which are must for cancer prediction. The symptoms are abnormal bleeding, cough, vomiting, change in bowel movement, unexplained weight loss. These symptoms often indicate cancer but it can be a reason for another disease. At the first stage, people often do not realize that they have cancer. It takes a long time to analyze that the patients are suffering from cancer.

Brain cancer has some symptoms which are very common for another disease. CT scan and MRI the major part which helps a doctor to identify that the patients have some abnormalities in their brain. After that Biopsy is the only medium to identify cancer disease. For identifying blood cancer blood test plays a vital role. Bone Marrow Aspiration also helps to identify blood cancer but the biopsy can help too sure about it. Ct scan, MRI and ultrasound helping in identify pancreatic cancer. Ultrasound, the digital rectal exam is the will tell any abnormalities in the prostate. A prostate-specific antigen is often present in blood it is not harmful but when it is higher level it can create a prostate infection, inflammation enlargement or cancer. For ovarian Cancer pelvic exam is a must. But sometimes it is failed to identify the small size of a tumor. To avoid that problem Transvaginal ultrasound (TVVS) which identify the only tumor. Pelvic MRI scan help to identify the more specific picture of a tumor. Only Biopsy can tell that cancer exists. The mammogram is the test of identifying breast cancer. But sometimes it is failed to identify cancer. Ultrasound test tells that the tumor is solid or liquid. MRI scan can help to find the affected area. Endoscopy checks irritation or abnormalities in the esophagus. Here generally analysis the symptoms and then also analysis some tests which are must for cancer prediction. Barium swallow test is performed to see the lining of the esophagus. Ct scan and MRI are used to identify a tumor in the lung. Sputum cytology test is used to look cancerous cell in the lung. Fecal occult blood test sometimes gives a wrong result for identifying colorectal cancer. A stool DNA test is more accurate to identify colorectal cancer. Flexible sigmoidoscopy detects polyps or cancer in the part of the colon. Colonoscopy is larger than sigmoidoscopy. It identifies a large number of spaces which is actually affected by cancer or not.

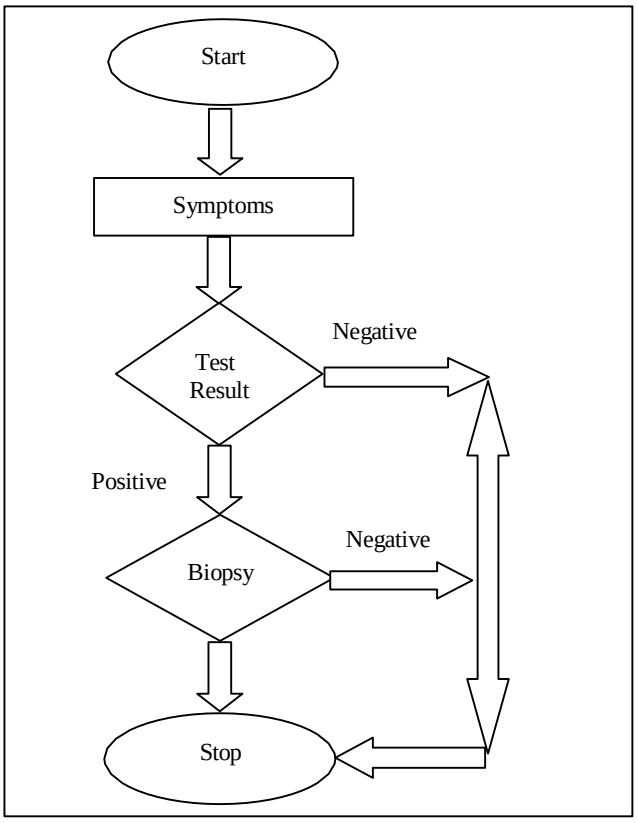

Fig 1: Dataflow Diagram of predicting cancer

In figure 1 first the process is get started with the start. Symptoms are indicated as the training part. If the symptoms are true then it showed the result. The test part is also working as the main part of predicting cancer. It the test part indicate true then the biopsy portion is performed otherwise the process is stopped. The test part and the biopsy both show negative then the process is stopped.

B. Data Mining Process

The data mining process works on a large number of datasets. It also predicts where the dataset will be. It is used for prediction, clustering, classification. There are two main section training and testing.

- Training is the part where the dataset is trained by input with expected output.
- Testing is a part where the evaluation of the model is shown in Fig 2.

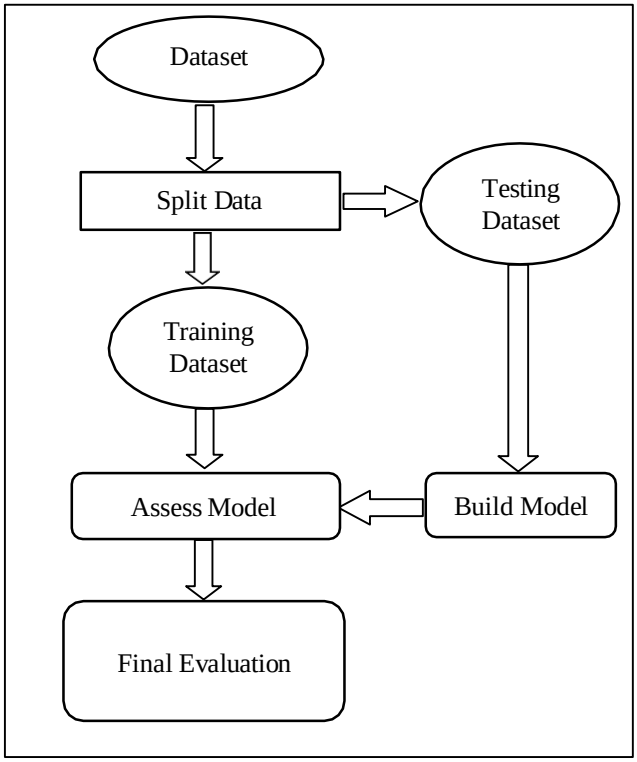

Fig 2: Data mining Diagram

VI. Result And Analysis

We research to predict cancer disease trying three types of algorithm and find the best accuracy among them. We use the Windows 10 operating system and Weka 3.6 version. Accuracy identifies the ability of classifier. The greater the accuracy will be a better classifier. So, our main work is to find the accuracy of all those three-classification algorithms. Among them, one will be greater in accuracy and that will be the best algorithm. We analyze 9 types of cancers accuracy, error rate, sensitivity, specificity, precision, F-score. Error rate finds the error of the dataset. Sensitivity finds actual true values and specificity finds actual negative values. The dataset will be ideal if FP=0, FN=0. Using 10-fold cross-validation and three classification learning algorithm Weka gives us a confusion matrix. Confusion matrix gives us the TP, FP, TN and FN values.

$$\text{Accuracy} = \frac{TP+TN}{P} + N \quad \text{.....(iii)}$$

$$\text{Error Rate} = \frac{FP+FN}{P} + N \quad \text{.....(iv)}$$

$$\text{Sensitivity} = \frac{TP}{P} \quad \text{..... (v)}$$

$$\text{Specificity} = \frac{TN}{N} \quad \text{..... (vi)}$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad \text{..... (vii)}$$

$$\text{F-score} = \frac{(2 * \text{precision} * \text{recall})}{(\text{precision} + \text{recall})} \quad \text{.....(viii)}$$

Here, TP=True Positive
TN=True Negative
FP=False Positive
FN=False Negative
P=Positive
N=Negative

TABLE III. Naive Bayes Using 10-Fold cross-validation

Class	*Accuracy*	*Error rate*	*Sensitivity*	*Specificity*	*Precision*	*F-Score*
Brain cancer	92%	8%	100%	91%	53%	69.2%
Blood Cancer	98%	2%	86%	100%	100%	92.5%
Pancreatic Cancer	98%	2%	86%	100%	100%	92.5%
Prostate Cancer	98%	2%	81%	100%	100%	89.5%
Ovarian cancer	100%	0%	100%	100%	100%	100%
Breast Cancer	98%	2%	88%	100%	100%	93.6%
Esophageal cancer	100%	0%	100%	100%	100%	100%
Lung cancer	99%	1%	94%	100%	100%	97%
Colorectal cancer	99%	1%	90%	100%	100%	95%
No cancer	100%	0%	100%	100%	100%	100%

After inserting dataset in Weka, we use 10-fold cross validation then we got a result. In the result window there is a confusion matrix. From the matrix, we got TP, TN, FP, FN, P, N values. We put those values in the proper equation and got the desired result for TABLE III. This TABLE III indicate the values of accuracy, error rate, sensitivity, specificity, precision and F-score using Naive Bayes classifier algorithm.

TABLE IV. K-Nearest Neighbor Using 10-Fold cross-validation

Class	*Accuracy*	*Error rate*	*Sensitivity*	*Specificity*	*Precision*	*F-Score*
Brain cancer	98%	2%	93%	99%	87%	90%
Blood Cancer	97%	3%	93%	98%	86%	89.4%
Pancreatic Cancer	98%	2%	90%	99%	92%	91%
Prostate Cancer	99%	1%	90%	99%	98%	94%
Ovarian cancer	100%	0%	100%	100%	100%	100%
Breast Cancer	98%	2%	92%	99%	92%	92%
Esophageal cancer	100%	0%	100%	100%	100%	100%
Lung cancer	99%	1%	94%	99%	97%	95.4%
Colorectal cancer	99%	1%	92%	99%	97%	94.4%
No cancer	100%	0%	100%	100%	100%	100%

After inserting dataset in Weka, we use 10-fold cross validation then we got a result. In the result window, there is a confusion matrix. From the matrix, we got TP, TN, FP, FN, P, N values. We put those values in the proper equation and got the desired result for TABLE IV. This TABLE IV indicate the values of accuracy, error rate, sensitivity,

specificity, precision and F-score using KNN classifier algorithm.

TABLE V. J48 Using 10-Fold cross-validation

Class	*Accuracy*	*Error rate*	*Sensitivity*	*Specificity*	*Precision*	*F-Score*
Brain cancer	98%	2%	92%	98%	90%	90%
Blood Cancer	97%	3%	94%	97%	82%	87%
Pancreatic Cancer	98%	2%	92%	99%	96%	93%
Prostate Cancer	98%	2%	89%	99%	97%	92%
Ovarian cancer	100%	0%	100%	100%	100%	100%
Breast Cancer	97%	3%	92%	98%	92%	92%
Esophageal cancer	100%	0%	100%	100%	100%	100%
Lung cancer	98%	2%	95%	99%	96%	95%
Colorectal cancer	99%	1%	90%	99%	98%	93%
No cancer	100%	0%	100%	100%	100%	100%

After inserting dataset in Weka, we use 10-fold cross validation then we got a result. In the result window, there is a confusion matrix. From the matrix, we got TP, TN, FP, FN, P, N values. We put those values in the proper equation and got the desired result for TABLE V. This TABLE V indicate the values of accuracy, error rate, sensitivity, specificity, precision and F-score using J48 classifier algorithm.

TABLE VI. Final Result

Class	*Accuracy*	*Error rate*	*Sensitivity*	*Specify*	*Precision*	*F-Score*
Naïve Bayes	98.2%	2%	92.5%	99.1%	95.3%	93%
KNN	98.8%	1.2%	94.4%	99.2%	94.9%	94.6%
J48	98.5%	2%	94.4%	98.9%	95.1%	94.2%

From TABLE VI it has shown the final analysis result. In this table is discussed about the average rate of accuracy, error rate, sensitivity, specificity, precision, F-score. From this, we could able to take a final decision. Here the average rate of accuracy of Naive Bayes is 98.2%. The average rate of KNN is 98.8% and for J48 it is 98.5%. The error rate of Naive Bayes is 2%. The error rate of KNN is 1.2% and for J48 it is 2%. The sensitivity of Naive Bayes is 92.5%. The sensitivity of KNN is 94.4% and for J48 it is 94.4%. The specificity of Naive Bayes is 99.1%. The specificity of KNN is 99.2% and for J48 it is 98.9%. The precision of Naive Bayes is 95.3%. The precision of KNN is 94.9% and for J48 it is 95.1%. The F-score of Naive Bayes is 93%. The F-score of KNN is 94.6% and for J48 it is 94.2%. From this, it is clear that KNN which is K-Nearest Neighbor gives an accurate result than the other two classifier algorithm. So, here K-Nearest Neighbor is performing best than other two classification algorithms.

Naive Bayes < J48 < KNN

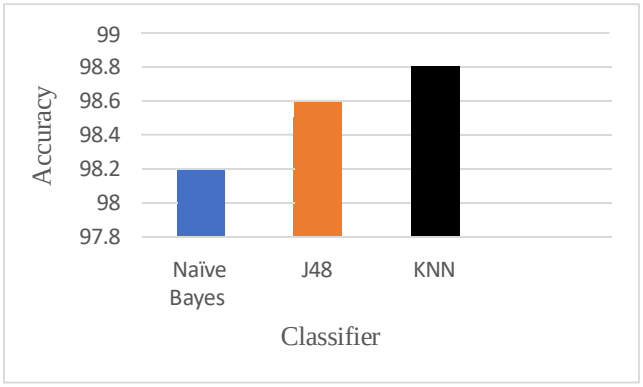

Fig 3: Graph of Accuracy

In Fig 3 the Black one represents K-Nearest Neighbor (KNN), the orange one represents J48 algorithm and the blue one represents Naive bayes Algorithms accuracy level.

VII. Conclusion

This paper is based on research medical dataset which able to predict Cancer disease. From the analysis, we can say that most of the cancer mainly happen for smoking, tobacco. Here three classification algorithms are applied to justify the dataset. In this purpose, we use the Weka tool. Weka is containing preparation, classification, clustering, regression, and visualization. The dataset is not creating any violence because it is not containing any personal data. It just contains some medical information. Three algorithms is used to identify confusion matrix. Confusion matrix gives the result of the classification algorithm. It contains information about actual & predicted classification. We have ten classes Brain cancer, Blood Cancer, Pancreatic Cancer, Prostate Cancer, Ovarian cancer, Breast Cancer Esophageal cancer, Lung cancer Colorectal, cancer, No cancer. It help to find accuracy, error rate, sensitivity, specificity, precision, and F-score. Tables are created for clear perception. It gives a comparison between three classification algorithm named Naive Bayes, K- Nearest Neighbor and j48 algorithm. The comparison table clearly declare which classification model is better .The three algorithm are works well but here K-Nearest Neighbor works more accurate than other two algorithm.

In the future, we will try to add more innovation to a large improvement. We will try to extend dataset. We will definitely try to use different preparation, classification, clustering, regression, and visualization. We will try to develop new models' prediction and survivability.

REFERENCES

1. Hong, W. K., et al. (2007). A Risk Model for Prediction of Lung Cancer. Journal of the National Cancer Institute, 99(9), 715–726.
2. Heuvelmans, M. A., et al. (2021). Lung cancer prediction by Deep Learning to identify benign lung nodules. Lung Cancer, 154, 1–4.
3. National cancer institute "What is Cancer "Available: https://www.cancer.gov/aboutcancer/understanding/what-is cancer [Accessed: 11.03.2019]
4. Quora "What is Classification in data mining?" Available: https://www.quora.com/What-is-classification-in-data-mining [Accesed:2-3-2019]
5. Quora "What is prediction in data mining?" Available: https://www.quora.com/What-is-prediction-in-data-mining [Accessed:02-03-2019]
6. Blog Aylien "Naïve Bayes" Available: http://blog.aylien.com/naive-bayes-for-dummies-a-simple- explanation [Accessed: 2.3.2019]
7. Wikipedia, the free encyclopedia "Naïve Bayes Classifier" Available: https://en.wikipedia.org/wiki/Naive_Bayes_classifier.