# Analysis and study on different cancer data sets –using classification methods of data mining tools

**# .PROF.KALAM NARREN**[1], Professor,Dept.of CSE,E
mail:Srinivas_research@yahoo.in **PROF.V.VINAY KRISHNA[2],** Asst.Professor,
Dept.ofCSE,Mail: itsmerams@gmail.com
# ChristuJyothi Institute of Technology and Science, Janagon, TS,INDIA

**Abstract**

The main aim of this paper is to check and compare classification ways with different (Lucamia )cancer knowledge sets victimization data processing tools to induce accuracy in medical results .Many researches have done and proved some results supported single knowledge sets and classifiers .we have examined internet based mostly knowledge sets and different knowledge sets for comparison. Classification results with neural network conjointly conferred here for our comparison**.**

## 1. INTRODUCTION

Data mining is that the method during which valuable data is extracted from the big dataset. it's reached the high growth over past few years. thanks to the quality of knowledge mining approaches in health world, it's become the nice technology in tending domain. This realization results in explosion of knowledge mining approaches [1]. Medical data processing will exploit the hidden patterns gift in voluminous medical knowledge that otherwise is left undiscovered. data processing techniques that are applied to medical knowledge embrace association rule mining for locating frequent patterns, prediction, classification and clump.

The analysis work drained data processing medical field is given as: Evans et. al [2] a system supported data processing techniques to notice the hereditary syndromes. Pradhan associate

d Prabhakaran [6] planned an approach through association rule mining to mine high-dimensional, statistic medical knowledge for locating high confidence patterns. DoronShalvi and St. Nicholas DeClaris, [4] mentioned medical data processing through unattended neural networks besides a way for knowledge image.

Traditionally data processing techniques were employed in varied domains. However, it's introduced comparatively late into the tending domain.

Normally the mandatory a part of any chassis is blood since it keeps one alive. It executes several important functions like to transfer O, carbonic acid gas, mineral and etc. to the entire body so as to stay metabolism. Blood consists of 3 main elements that RBC, white blood corpuscle and Platelets. insufficient quantity of the blood may have an effect on the metabolism critically that can be terribly venturesome if early treatment isn't taken. one amongst the traditional blood disorders is malignant neoplastic disease. malignant neoplastic disease is that the common form of cancer in

youngsters. All cancers begin in body cells, and malignant neoplastic disease may be a form of cance cells .

## 2. Related Work

Many researches concentrate on the cancer diseases. The regression toward the mean algorithmic program applied on the Blood cancer dataset considering varied demographic and clinical characteristics of patients [9]. Authors have compared the various classifiers call tree, Multi-Layer Perception, Naive mathematician, ordered stripped-down improvement, and Instance based mostly for K-Nearest neighbor on 3 totally different databases of carcinoma by mistreatment classification accuracy and confusion matrix supported 10-fold cross validation technique in WEKA[5]. Authornames et al finished the support vector machine as best classifier in terms of accuracy for prediction of carcinoma that is that the sixth commonest cancer and a significant health problem within the world [6].

The studies disclosed that reckoning on the sort of dataset used every model differs in their performance. If thedataset consists of unlabeledfeatures, then the clump model higher suits for pattern recognition among the many strategies k-means algorithmic program [7].

Some studies centered on heart diseases and analyze dataset mistreatment NaveBayes, K-NN, and call List algorithmic program mistreatment Tanagra data processing tool[10].The study to analyze comparison of seven totally different classification algorithms particularly, Naive mathematician, Naive mathematician updatable, FT Tree, KStar, J48, LMT, and Neural network for analyzing infectious disease prognostic knowledge has been conferred. The study concludes thatthe Naive mathematician classification performance is best than different classification techniques for infectious disease dataset . By analyzing varied techniques through analysis papers, this paper has elect the foremost correct technique for prediction of primary tumors. Next section discusses the proposed technique to be used.

### 2.1 BACKGROUND OF LEUKEMIA

Leukemia may be a form of cancer of the blood or bone marrow categorise by associate degree irregular augment of undeveloped white blood cells referred to as "blasts." it's a thick term covering a compilation of diseases. per yank Cancer Society it's approximated that 48,610 persons (27,880 men and twenty,730 women) are notice with and twenty three,720 men and girls can terminate of malignant neoplastic disease in 2013 solely. In turn, it's a part of the even broader set of diseases heavy the blood, bone marrow, and bodily fluid system, that are all referred to as hematologic growth. Over time, malignant neoplastic disease cells will force out the traditional blood cells. this will cause serious issues like anemia, bleeding, and infections. malignant neoplastic disease cells also can unfold to the liquid body substance nodes or different organs and cause swelling or pain. There are many differing types of malignant neoplastic disease.

. Acute lymphocytic leukemia, or ALL.

. Acute myelogenous malignant neoplastic disease, or AML.

. Chronic cancer of the blood, or CLL.

. Chronic myelogenous malignant neoplastic disease, or CML.

In general, malignant neoplastic disease is classified by how briskly it gets worse and what reasona cell it affects. Acute {lymphoblastic malignant neoplastic disease|lymphocytic leukemia} (ALL) is that the most general-purpose form of leukemia in young youngsters and Acute Myelogenous malignant neoplastic disease (AML) happens additional sometimes in adults than in youngsters, and additional sometimes in men than girls [12]. The young white blood corpuscle also can build up in a very form of extreme dullard sites, particularly the mining's, gonads, thymus, liver, spleen, and liquid body substance nodes. Therefore thanks to extreme lye- psychoneurotic blast or myeloid blast within the marrow they additionally low into the peripheral blood stream. Acute chronic leukemia (AML) is additionally recognized by different names, that embrace acute granulocytic leukemia, acute Myelogenous malignant neoplastic disease, acute leucaemia, and acute non-lymphocytic malignant neoplastic disease. "Acute" means this malignant neoplastic disease will develop speedily if not treated, and would just about definitely be deadly in a very few months. "Myeloid" refers to the sort of cell from wherever this malignant neoplastic disease begins. In most cases AML build up from cells that will wind into white blood cells (other than lymphocytes), however in some cases of AML expand in different styles of blood forming cells.

AML starts within the bone marrow (which is that the soft inner a part of bound bones, wherever new blood cells are fashioned), however in most cases it apace moves into the blood. It will typically widen to different components of the body along with the liquid body substance nodes, liver, irritate inner system (brain and spinal cord), and testicles. different styles of growth will begin in these organs and behind that expand to the bone marrow. however these cancers that begin anyplace else so increase to the bone heart aren't leukemia's. [13]

Diagnosing malignant neoplastic disease relies on the actual fact that white corpuscle count is bigger than before with immature blast (lymphoid or myeloid) cells and attenuated neutrophils and platelets. The group action of excess range of blast cells in marginal blood may be a significant symptom of malignant neoplastic disease. therefore hematologists habitually examine blood smear below magnifier for correct identification and classification of blast cells [14].

2 .1.1 **Causes and Risk Factors of malignant neoplastic disease**

The satisfactory causes of malignant neoplastic disease are unidentified and in most case its unsettled why malignant neoplastic disease has developed. analysis into potential causes goes on all the time. Like different cancers, malignant neoplastic disease isn't transferable and can't be approved on to others. There are many of things which will amplify a person's risk of budding malignant neoplastic disease. Having a scrupulous hazard issue doesn't denote you'll positively get this class of illness and personnel lacking any recognized risk factors will still develop it. The recognized risk issue of generate this kind of cancer i.e. malignant neoplastic disease are clarify here.

- Exposure to radiation: those that exposed to high level of unleash, like nuclear developed accidents, have a main risk of developing malignant neoplastic disease than those that haven't been exposed. On the opposite hand, alittle numeral of individuals within the UK are uncovered to emission levels high adequate augment their risk.
- Smoking: Smoking increase the chance of initial malignant neoplastic disease. this might ensue to the extraordinary levels of aromatic hydrocarbon in cigaret smoke.
- Exposure to benzene: In terribly uncommon cases, malignant neoplastic disease might begin thanks to the long run contact to aromatic hydrocarbon (and presumably different solvents) employed in trade.

- Cancer treatments: currently so, some anti-cancer treatments like therapy or therapy malignant neoplastic disease to create up once some years of this behavior. the chance increase once persuaded styles of therapy medicine are mutual with therapy. whereas malignant neoplastic disease develops since of earlier anti-cancer treatment, this is often referred to as lower malignant neoplastic disease or treatment connected malignant neoplastic disease.
- Blood disorders: folks with bound blood disorders, like myelodysplasia or myeloproliferative disorders have a distended risk of initial AML.
- Genetic disorders: folks with bound hereditary disorder, excluding Down's syndrome and Franconia's anemia, have associate degree inflated risk of embryonic malignant neoplastic disease.

Other less general symptoms could also be caused by a rise of malignant neoplastic disease cells in a very finical space of the body. Your bones might ache, reason by the strain from a buildup of undeveloped cells within the bone marrow. you would possibly additionally distinguish raised, light-blue wine areas below the covering thanks to malignant neoplastic disease cells within the skin, or swollen gums caused by malignant neoplastic disease cells within the gums. [12] Cancer starts once cells in a very piece of the body begin to rise out of management and might be different areas of the body.

## 3. PROPOSED FRAMEWORK

Some clarification needed

***Methodology***

Process 1: begin from uploading the samples that is within the variety of info file. we tend to choose the info file and transfer it into the interface homepage

Process 2: The uploaded file contains the cistron data regarding the patients United Nations agency have malignant neoplastic disease

Process 3: once knowledge has been uploaded ,create a graphical illustration of knowledge for image

Process 4: currently we tend to applied the Genetic algorithmic program for the reduction of cistron set

Process 5: For reduction of cistron set we tend to calculate the subsequent G.A steps

where are The GA steps

Process 6: Calculate total cistron data as population

Process 7: Apply fitness operate on cistron knowledge

Process 8: choose sub population and appraise a brand new fitness operate

Process 9: Finally appraise best fitness operate

Process ten : knowledge set is reduced referred to as reduced cistron set

Process eleven : finally, classification is finished by mistreatment neural network that Back propagation is employed.

Process twelve : Matched and unmatched values are categorised

Process 13: Compare these values matched and unmatched and appraise the accuracy by mistreatment accuracy parameters

Process 14: Accuracy parameters are fault acceptance rate (FAR) and fault rejection rate (FRR) is calculated

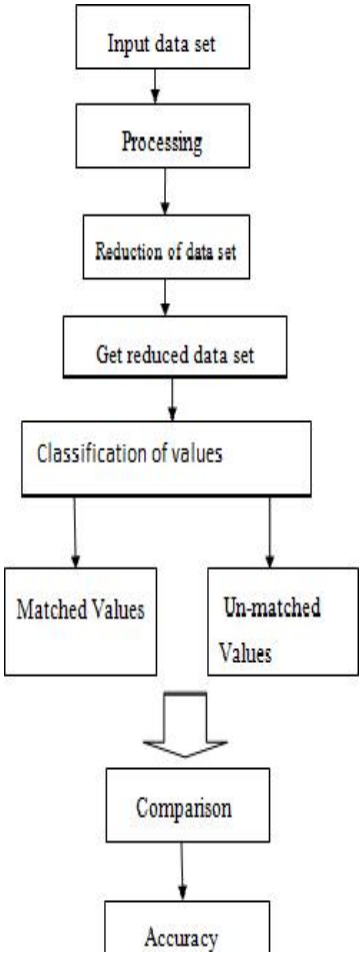


**Fig 3.1: Flow chart of methodology**

**2.2**

## 2.3 4. RESULTS & IMPLEMENTATION

**Fig 3.2: GUI interface**

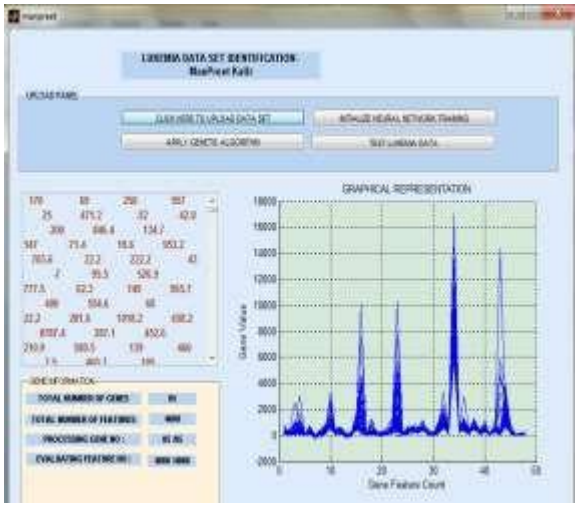


**Fig 3.3: GUI up load the interface**

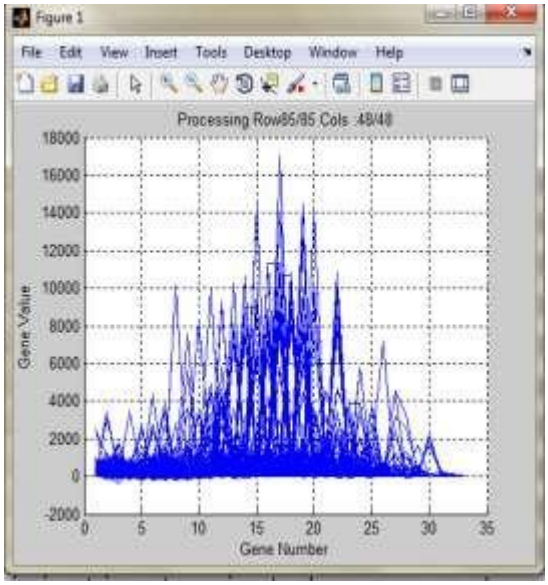


**Fig 3.4: Genetic algorithm graph**

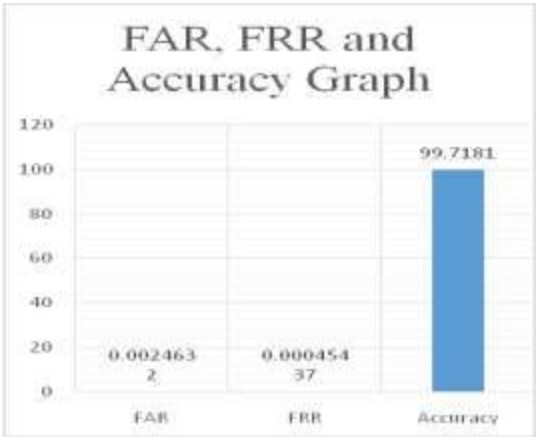


**Fig 3.5: FAR, FRR and Accuracy Graph**

The higher than figures shows the graphical program panel of the planned system having totally different guicontrols of the user panel having transfer buttons to transfer the information set, Genetic algorithmic program and neural network intialization and testing button.This panel contains push buttons like click here to transfer knowledge set ,apply genetic algorithms,intialize neural network and check leukemia knowledge .At left bottom,code data is made.

It shows the information uploading method and therefore the genes graphical illustration with their process rows and columns to the full range of rows and columns. The graphical illustration is between the every feature count of cistron worth. the information set values are uploaded within the edit button of interface as shows within the higher than figure.Code data box incorporates total range of genes, total range of options, process cistron range, evaluating feature no.This graph represents the cistron worth and cistron feature count per this graph massive cistron worth having best fitness operate. Graphical illustration of cistron knowledge set provides image of knowledge sets .

## 5. CONCLUSION & FUTURE SCOPE

Leukemia may be a cancer of the marrow and blood. This adversely affects the formation and traditional operate of blood tissues and cells. knowledge sets are notably enticing for diagnostic troubles while not a linear resolution. sometimes physicians analyze clinical and laboratory symptoms of blood cancer qualitatively and at last use bone marrow biopsies as a more robust procedure for assess the character of illness. The precise and reliable detection of cancer want additional Para clinical tests and prices and take abundant time. during this investigation we tend to apply easy and early clinical and assessment for correct detection of malignant neoplastic disease. so we will use trained ANN and Genetic Algorithms additionally ,so that we will predict cancer with least clinical and laboratory tests and while not obligation of abundant time. Accuracy of the detection of cancer by the assembled artificial neural network was analyze by mythical creature and multivariate analysis. Outputs of trained (classification algorithms) for testing knowledge were wont to plot graphs.

The future work of our experimentation can embrace ever-increasing the amount of records of {the knowledge|theinfo|theinformation}set for coaching the data so as to induce correct results and therefore the network are able to learn additional professionally ,with additional range of records. additionally to the present, larger datasets may be obtained &applied; and therefore the approach may be tested to supply higher accuracy results. totally different neural networks and different classification techniques also can be tried to get higher results. Use of support vector machines are thought of within the future work as a classification tool.

## 6. References:


1. Li, Eldon Y. "Artificial neural networks and their business applications."Information& Management 27.5 (1994): 303-313

2. Xue-wen Chen and Michael McKee," Finding expressed genes using genetic algorithms and support vector machines", Department of Electrical and Computer Engineering, California State University 18111 Nordhoff Street, Northridge, CA 91330, USA,2003
3. Marc C. Chamberlain, M.D.," Leukemia and the Nervous System",2003

4. S. H. Rezatofighi et.al," A New Approach to White Blood Cell Nucleus Segmentation Based on Gram-Schmidt Orthogonalization", InternationalConference on Digital Image Processing,2005.

5. Gouda I. Salama et.al, "Fuzzy Analysis of Breast Cancer Disease using Fuzzy c-means and Pattern Recognition", *Southeast Europe journal of soft computing*,2011

6. Li-Yeh Chuang et.al, "Support Vector Machinebased Prediction for Oral Cancer Using Four SNPs in DNA Repair Genes", *Proceedings of the International Multi Conference of Engineers and Computer Scientists 2013*Vol. I, 2011.

7. K.Lokanayaki et.al, "Exploring on Various Prediction Model in Data Mining Techniques for Disease Diagnosis", *International Journal of Computer Appli*cations (0975 – 8887), Vol. 77, No.5, September 2013
8. Alba, Enrique, et al. "Gene selection in cancer classification using PSO/SVM and GA/SVM hybrid algorithms." Evolutionary Computation, 2007. CEC 2007. IEEE Congress on. IEEE, 2007.
9. NiponTheera-Umpon," Morphological Granulometric Features of Nucleus in Automatic Bone Marrow White Blood Cell Classification", IEEE transactions on information technology in biomedicine, VOL. 11, NO. 3, MAY 2007.

10. YvanSaeys and et.al," A review of feature selection techniques in bioinformatics", Vol. 23 2507–2517
11. Raje, Chaitali, and JyotiRangole. "Detection of Leukemia in microscopic images using image processing." Communications and Signal Processing (ICCSP), 2014 International Conference on. IEEE, 2014
12. El-Nasser, Ahmed Abd, Mohamed Shaheen, and Hesham El-Deeb. "Enhanced leukemia cancer classifier algorithm." Science and Information Conference (SAI), 2014. IEEE, 2014.
13. Zadeh, HosseinGhayoumi, SiamakJanianpour, and JavadHaddadnia. "Recognition and Classification of the Cancer Cells by Using Image Processing and Lab VIEW." International Journal of Computer Theory and Engineering(2013).
14. Mohapatra, Subrajeet, et al. "Fuzzy based blood image segmentation for automated leukemia detection." Devices and Communications (ICDeCom), 2011 International Conference on. IEEE, 2011