Annals of R.S.C.B., ISSN:1583-6258, Vol. 24, Issue 2, 2020, Pages. 748- 763
Received 24 October 2020; Accepted 15 December 2020

A Study on Various Machine Learning Techniques Used For Colorectal Cancer Disease Prediction and Survival

BalajiVicharapu[1]*, AnuradhaChinta[2] and S.R. Chandra Murty Patnala[3]

[1]Research Scholar, Department of CSE, AcharyaNagarjuna University, A.P, India

[2] Assistant professor, Department of CSE, V R Siddhartha Engineering College, India

[3] Research Supervisor, Department of CSE, AcharyaNagarjuna University, A.P India

ABSTRACT

Cancer is a major burden of disease worldwide. Each year, tens of millions of people are diagnosed with cancer around the world, and more than half of the patients eventually die from itCancer is a disease caused by the uncontrolled division of abnormal cells in parts of the body. There are several types of cancer in the world. Inthepast, the diagnosis of cancer mainly depended on the experience and knowledge of the doctor. Butbecausein thepastten years,computerized decision support systems haveplayed a vital role in the healthcare industry. Machine learning methods are used forthe prevention and early detection of cancer patients. The rapid development of machinelearningtechnology does help cliniciansmakecorrect diagnosis decisions. Tumor prediction at the TNM (tumor, nodule, and metastasis) stage of colon cancer has been studied using the most influential histopathological parameters and five-yeardisease-free survival (DFS) predictionthrough machine learning (ML) clinical research.4021 patients wereselectedfor analysis from the ColonCancerRegistry (CRC)at Chang Gung Memorial Hospital, Linkou, Taiwan.Several ML algorithms were usedto predict the tumor stage of colon cancer takinginto account thetumoraggressionscore(TAS).The performance of the various ML algorithms was assessed using five-foldcross-validation, which resultedin an efficient validation of the precision achieved by the algorithms, whichtook intoaccount boththe standard TNM staging cases andthe TNM staging with the tumoraggressionscoreThe random forestmodel was observed to achieve an F-value of 0.89 when tumor aggression was included as the attribute score along with the standard attributes normally used for TNM stage prediction. We also found that the Random Forest algorithm outperformed all other algorithms, with an accuracy of about 84% and an area under the curve (AUC) of 0.82 ± 0.10,around the five-year DFS.

Keywords: colon cancer; artificial intelligence; machine learning; Supervised Learning; Support Vector Machines ; prediction

I. Introduction

Cancer is one of the most common diseases in the world.Because of the rampant growth of cells in any tissue or body part, a large number ofdeaths are caused. Cancer is a fatal disease caused bychanges in normal cells in the body.In the body,the cells thatcause tumors grow uncontrollably, except for leukemia,which is the main cause of cancer. Ifthe tumor is not treated in time, it will grow and spread through the bloodstream tothe surrounding area, affecting the digestive system, resources and health. Men are more susceptible to lung, prostate, stomach, and liver cancer, while women are more susceptible to breast, colon, lung, cervical, and stomach cancers.A related disease is called cancer. If the correct treatment and diagnosis of the disease is not timely(this is thecase in most cases), this malignant disease can even leadto death.

According to WHO (World Health Organization) data, 8.2 million people die from cancer eachyear. It is estimated that 13% of alldeaths worldwide are caused by cancer. It is estimatedthatnew cancercaseswill increaseby 70% in the next 20years. Thereare 100 types of cancer, each of which requiresa uniquediagnosis and treatment. The most commonly diagnosed cancers in the world are lung cancer (1.8 million, 13% of the total),breast cancer (1.7 million, 12% of the total),and colon cancer (1.4 million, 9.7 of the total)). The most common cause of cancer death islung cancer (1.6 million, 19th1% of the total)liver cancer (800,000, 9.1% of the total)gastric cancer (700,000,8.8% of the total). It is estimated that by

Annals of R.S.C.B., ISSN:1583-6258, Vol. 24, Issue 2, 2020, Pages. 748- 763
Received 24 October 2020; Accepted 15 December 2020

2025,due to population growth and aging, as many as 19.3 million new cancer cases may be addedeachyear. Cancer is one of the maincauses of death in countries around the world [1]. Figure 1 show [2] the approximatenumber of the most common cancers in India and shows the severity of the problem.

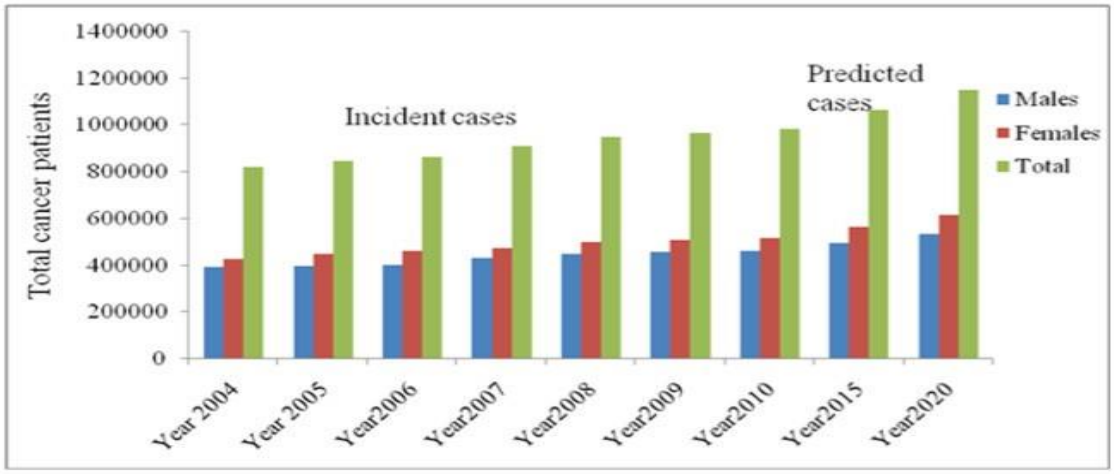


Figure 1: Cancer Statistics Scenario in India

Colon cancer is the third most common cancer in the world,withmore than 1.8 million new patients in 2018 [1]. According to a report from the Taiwan Health Promotion Agency,colon cancer wasamong the most common cancers in 2016 [2]. Intermsof mortality, colorectal cancer is the second leading cause of death in the United States.Accordingto the2012-2016 mortality statistics, thereare14.2 deaths per 100,000 men and women eachyear. [3]. Colon cancer was the third leading cause of cancer deaths inTaiwan in 2016.In2016, 47,760peopleidiedof cancer, accounting for 27.7% of the total deaths. The crude death rate of CRC is 24.3 deaths per 100,000 people. ,Increaseby 0.1% Comparedwith the previous year; the age-adjusted death rate was 14.6 deaths per 100,000, a decrease of 0.3%, which resulted in 12% of all cancer deaths in 2016. In addition, it was also found that the mortality rateof men (28.1 per 100,000 people)is higher thanthatof women (20.6 per 100,000 people) [4]. A study by the American Cancer Society [5] predicts that duetolifestylechanges, people aged20 to 34 and 35 to 49 yearsold have a 90% and 27.7% increasein the risk of colorectal cancer.

For each colon cancerpatient,collect different parameters, such as gender, age, body mass index (BMI), family history (HF),smoking, drinking, and pathological data, suchas tumor size, tumor differentiation, peripherallesions, tumor staging,and knots. The staging of festivals, etc. In severalstudies, tumor size is an important factor in determining the staging of colon cancer.

II.RELATEDWORK

Inorder to discover different methods for diagnosing different types ofcancer, a lot of work and research have beencarriedout. Itattempts to predict and diagnose cancer based on symptoms that appear early. Thisarticle [3] focuseson lung cancer, which is a fatal lung disease. On this basis, 20 parameters arelistedin the feature selection process. Some influencing parameters are weight loss, blood mucus, back pain andsoon.The most

importantstepis tounderstand the patients whocan accept specialdiagnosticprocedures; theyfound that the form of supervised learning ismuch better than the cross-validationmethod; from this study,they found that random tree classifiers,artificial neural networks,logistics,hierarchical perception andMinimalorder optimizationperforms better and more reliable in this regard. Thestudy in [4] used artificial neural networks(ANN), logistic regression, and naive Bayes methodsto study breast cancer and disease prognosis. Thepurpose of this study is to provide the following results;on the one hand, it evaluates the grammatical quality of medical data sets, and on the other hand, it evaluates the applicability of data extraction techniques todata. Finally, by applying artificial neural networks (ANN), logistic regression and naive Bayes, the knowledge extracted from the data set is used to predict diseases. Ithasbeen found that these methodshavethe highest lift coefficients for most class values. This work uses cancerdiagnosticmethods. Severalwell-known classification algorithmsare used to diagnose cancer, such as DT (Decision Tree), SVM (Support Vector Machine), KNN (KNearest Neighbor) and NN (Neural Network).Somepeoplebelieve that the classification process depends on the importance of various features in the collected data. The author of [5] found that medical disease data usually containsbothpoisoning data and limit data. Noisy data. To improveaccuracy, they used an ant colony optimizationmethod.In their work,the authors of [6] useddifferent data mining methods, such as classification, mining classificationrules, soft computingmethods,neuralnetworks, and fuzzy logic todiagnose oral cancer. Theypointedout the effectiveness of each of the previousmethods for classification problems in the medicalfield,andproved the importance of genetic algorithms in optimizing data mining algorithms toimprove prediction accuracy.The author checked the applicability of Apriori and decisiontrees to discover important patterns ofdiscovery. The goal is to develop a lung cancer prediction system based on meaningful patterns. The prediction system can detect a person’s propensity for lung cancer, 400 datesfrom cancer patients and no cancer. Using the suggested method, the author indicated that a statistically significant association can be found from the collecteddata,but the results evaluated using the suggested method did not show high statistical reliability.

III. MACHINELEARNING

Machine learning (ML) can be interpreted as automating and improving the learning process of a computer based on yourexperience, without the need for actual programming without human assistance. The process starts by transmittinghigh-qualitydata, and then training our machines (computers) by using the data and various algorithms to create machine learning models.Thegoal of machine learning is to learn from data. There have been many studies on how to let the machine learn by itself [2] [3]. Many mathematicians and programmers use different methods to find solutionsto this problem. Some of them are shown in Figure 2

Annals of R.S.C.B., ISSN:1583-6258, Vol. 24, Issue 2, 2020, Pages. 748- 763
Received 24 October 2020; Accepted 15 December 2020

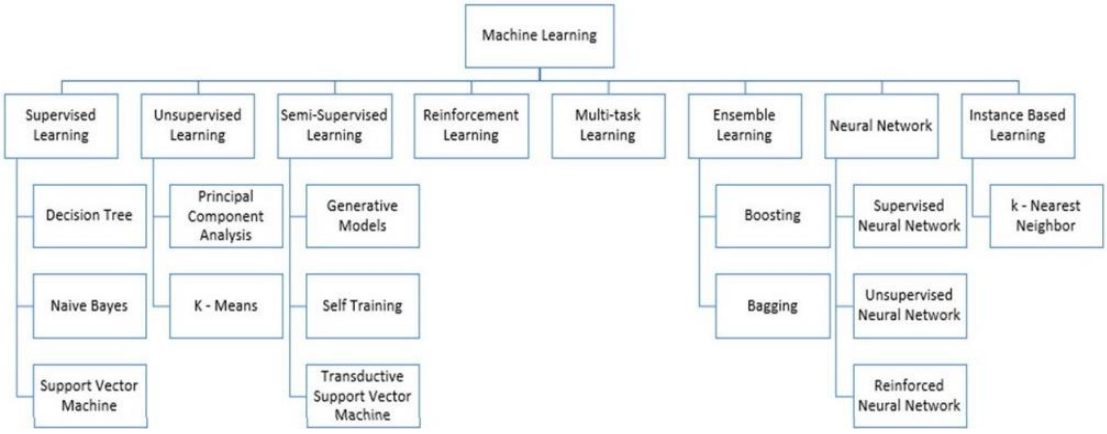


Figure 2: Types of Machine Learning

We used different machine learning classifiers, such as random forest, support vector machine, logistic regression, multi-layer perceptron, K-Nearest Neighbors, and Adaptive Boosting.; these classifiers are trained using supervised learning and use various metrics such as accuracy , Accuracy,memory and F measure) to evaluate theperformance of the model. Preprocess the collected data by removing missing values, and normalize the rest of the data**.**

All parameters are obtained from various sources in the clinical system, as shown in Table 1.

Table 1. Sources of parameters collected in the clinical system. TNM: tumor, node, and metastasis.

Sources	Parameters Collected
Chart records	Age, gender, adjuvant therapy, status of follow-up, medical illness, pre-operation lab data
History taking	Smoking history, coffee consumption, alcohol consumption, physical activity
Intra-operative finding	Operation date, intent of resection, operation timing, operation finding, operation type, early morbidity, late morbidity, mortality
Histo-pathologyReports	Tumor location, gross appearance, circumferential involvement, tumor size, histologic type, histologic grade, tumor extension, examined lymph node number, total positive lymph node number, TNM staging

As presented in Table 2, the parameters included in the study were: body mass index (BMI) that was categorized as <18.5, underweight; 18.5–23.9, normal; 24.0–26.9, pre-obesity; and exceeding 27 Causes obesity; family history, describing whether the patient's immediate family has cancer or other geneticdiseases; age classificationisandlt; fifty; and 50; thepatient also has a history of hypertension, diabetes, and personal habits such as smoking and drinking.), hemoglobin level (abnormal: and<11g/dl and normal: ≥11 g/dl), albumin level (LAB_ALB) (normal and<3.5g/dl and abnormal ≥3.5 g/dl),creatinine level (LAB_CR) And white blood cell count (WBC),these parameters are collected using preoperativeprocedures and blood samples obtained during patient follow-up. The patient's histopathological parameters were also recorded, such as tumor length (Tumrlen), tumor width (Tumrwid), peripherallesions (CirInvo), tumor differentiation (TumrDiff), tumor stage (T stage), and lymph node stage (N stage). ... Thesehistopathological parameters are collected from patients after biopsy or surgery. Comparedwith other stages, the data includes more patients who belong to tumor stage T3, as shown in Figure 2, where 1,2,3, and 4 are indicated,representing tumor stages T1, T2, T3, and T4, respectively.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 24, Issue 2, 2020, Pages. 748- 763
Received 24 October 2020; Accepted 15 December 2020

Table 2. Parameters used in this study with *p*-value.

Parameters	Tumor Aggression Score		*p*-
	<9.8 (3709)	≥9.8 (294)	
BMI			0.004
<18.5	215 (5.8)	35 (11.90)	
18.5–23.9	1665 (44.89)	151 (51.36)	
24.0–26.9	1070 (28.85)	66 (22.45)	
≥27	759 (20.46)	42 (14.29)	
Family History (FH)			<0.001
No	2145 (57.83)	180 (61.23)	
Yes	1429 (38.53)	104 (35.37)	
Unknown	135 (3.64)	10 (3.4)	
Age			0.007
<50	527 (14.20)	50 (17)	
≥50	3182 (85.80)	244 (83)	
Gender			<0.001
Male	2114 (57)	165 (56.12)	
Female	1595 (43)	129 (43.88)	
Hypertension			<0.001
Yes	2447 (65.97)	191 (64.96)	
No	1262 (34.03)	103 (35.04)	
Diabetes			<0.001
Yes	3136 (84.55)	231 (78.57)	
No	573 (15.45)	63 (21.43)	
Smoking			0.001
Never	2324 (62.66)	174 (59.18)	
Ex-Smoker	546 (14.72)	42 (14.29)	
Current	839 (22.62)	78 (26.53)	
Alcohol			<0.001
Never	2622 (70.69)	213 (72.45)	
Ex-Drinker	218 (5.88)	18 (6.12)	
Current	869 (23.43)	63 (21.43)	
CEA Level			<0.001
<5	2424 (65.35)	145 (49.32)	
≥5	1285 (34.65)	149 (50.68)	
Hemoglobin			0.9
Low (<11)	853 (23)	182 (61.90)	
Normal	2856 (77)	112 (38.10)	
LAB_ALB			<0.001
≤3.5	424 (11.43)	128 (43.54)	
>3.5	3285 (88.57)	166 (56.46)	
LAB_CR			<0.001
≤1.1	2954 (79.64)	233 (79.25)	
>1.1	755 (20.36)	61 (20.75)	
WBC			<0.001
≤5500	202 (5.5)	14(4.8)	
>5500	3507 (94.5)	280(95.2)	
OP Time			0.001
Elective	3635 (98)	284 (96.6)	
Emergency	74 (2)	10 (3.4)	
OP Find			<0.001
None	3199 (86.25)	205 (69.73)	
Combined	470 (12.67)	84 (28.57)	
Any one	40 (1.08)	5 (1.7)	
CirInvo			<0.001
No	1972 (53.17)	26 (8.84)	
Yes	1737 (46.83)	268 (91.16)	
Tumor Differentiation			<0.001
Grade I	477 (12.86)	7 (2.38)	
Grade II	3001 (80.91)	183 (62.24)	
Grade III	231 (6.22)	104 (35.37)	
Tumor Width			<0.001
≤4.4	2582 (69.61)	8 (2.73)	
>4.4	1127 (30.39)	286 (97.27)	
Tumor Length			<0.001
≤4.4	2679 (72.22)	10 (3.4)	
>4.4	1030 (27.78)	284 (96.6)	
T stage			<0.001
T1	377 (10.16)	5 (1.70)	
T2	531 (14.32)	4 (1.36)	
T3	2322 (62.61)	184 (62.59)	
T4	479 (12.91)	101 (34.35)	
N stage			<0.001
N0	2062 (55.6)	179 (60.89)	
N1	1010 (27.23)	57 (19.39)	
N2	522 (14.07)	46 (1.24)	
N3	115 (3.10)	12 (4.08)	

There are three main learning styles or learning models that an algorithmfollows.

i).s**upervised learning:**

The input is called training data and has a label or result calledspam/non-spam or one stock price at a time. The model is built inthe training process thatrequires predictions and corrects when these predictions are incorrect. The training process continues until the model reachesthe accuracy requiredby the training data. Examples ofproblems: classification and regression. Examples of algorithms: logisticregression and neuralback propagationnetworks.

ii).**Unsupervised Learning:**

Theentry is unlabeled and thereareno knownresults.The model is created by outputting the structure that exists in the input data. Thiscan be a rule of thumb for extraction.Systematicallyreducing redundancy or organizing data based on similarity can be a mathematical process. Examplesof problems are clustering, dimensionality reduction, and learning associationrules. Examplesof algorithms include: apriori algorithm and KMeans.

iii).**Semi-Supervised Learning:**

Annals of R.S.C.B., ISSN:1583-6258, Vol. 24, Issue 2, 2020, Pages. 748- 763
Received 24 October 2020; Accepted 15 December 2020

The input data is a mixture of labeled and unlabeled examples. There are prediction problemsyou want, but the model needs to learn the structure to organize the data and make predictions. Examplesof problems are classification and regression. The example algorithm isan extension of other agile methods that make assumptions about modeling unlabeled data.

IV. MACHINE LEARNING ALGORITHMSGROUPED BYSIMILARITY

S.No	Algorithm	Description	Example Algorithms
1	Regression Algorithms	Regression models the relationship between variables and uses the error metric in the model predictionto refineititeratively. Regression methods are the mainforce of statistics and have been integrated	i. Ordinary LeastSquares ii. Regression(OLSR) iii. LinearRegression iv. LogisticRegression v. StepwiseRegression
2	Instance-based Algorithms	Instance-based learning is a decision-making problem thatcontains training data instances or examples that are deemed important or necessaryfor the model	i.k-Nearest Neighbor (KNN) ii.Learning Vector Quantization (LVQ) iii.Self-Organizing Map (SOM) iv.Locally Weighted Learning (LWL)
3	Regularization Algorithms	An extension of another method (usuallya regression method) penalizes the model due to its complexity andfavorsa simpler model that is moresuitablefor	i.RidgeRegression ii.Least Absolute Shrinkage and Selection Operator (LASSO) iii.ElasticNet iv.Least-Angle Regression
4	Decision Tree Algorithms	Decisions fork in tree structure before prediction.Make a decision on thisitem. Decision trees are trained using data from classification and regression problems. Decision trees are usually fast and accurate, and are very popular in machine learning	i.Classification and Regression Tree(CART) ii.Iterative Dichotomiser 3 (ID3) iii.C4.5 and C5.0 (different versions of a powerful approach) iv.Chi-squared Automatic Interaction Detection (CHAID) v.DecisionStump
5	Bayesian Algorithms	Bayesian methods are those that explicitly apply Bayes' Theorem for problems such as classification and regression	i.Naive Bayes ii.Gaussian Naive Bayes iii.Multinomial Naive Bayes iv.Averaged One-Dependence Estimators(AODE)
6	Clustering Algorithms	Clustering methods are typically organized by the modeling approaches such as centroid-based and hierarchal. All methods are concerned with using the	i.k-Means ii.k-Medians iii ExpectationMaximisation (EM)

7	Artificial Neural Network Algorithms	ArtificialNeuralNetworksaremodels that are inspired by thestructureand/or function ofbiological neural networks. They area class of pattern matching that arecommonly used for regression andclassification problems but are reallyan enormous subfield comprised ofhundreds ofalgorithmsandvariations for all manner of problemtypes.	i.Perceptron ii.MultilayerPerceptrons (MLP) iii.Back-Propagation iv.Stochastic Gradient Descent v.Hopfield Network vi.RadialBasis FunctionNetwork (RBFN)
8	Dimensionality Reduction Algorithms	Clustering methods, dimensionalityeductionseekandexploittheinherentstructure in the data, but inthis case in an unsupervised manneror order to summarize or describedata using less information. This canbe useful to visualize dimensionaldata or to simplify data which canthen be used in a supervised learningmethod.	i.Principal ComponentAnalysis (PCA) ii.PrincipalComponentRegression (PCR) iii.PartialLeastSquaresRegression (PLSR) iv.MultidimensionalScaling(MDS) v.Linear DiscriminantAnalysis (LDA)
9	Ensemble Algorithms	Ensemblemethods are models composed of multiple weaker models thatare independently trained and whosepredictions arecombined in some way to make theoverall prediction.	i.Boosting ii.BootstrappedAggregation (Bagging) iii.AdaBoost iv.StackedGeneralization (Stacking) v.Gradient Boosting Machines (GBM) vi.Gradient Boosted Regression Trees (GBRT) vii.Random Forest

V.MACHINE LEARNING APPLICATIONS IN HEALTHCARE

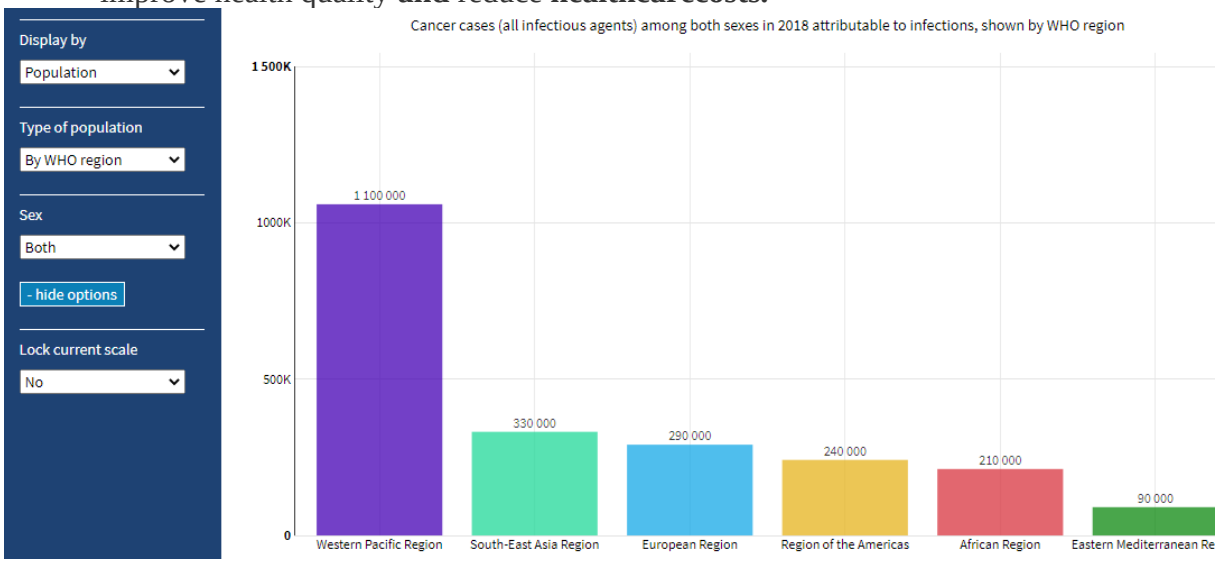
To **make** the **right** decisions in **healthcare,machinelearning** is **essential. Thevarious applications** are shown **below.**

- **Through** training, chronic **diseasescan be identified, and** patients **can be prioritized according to their complications so** that they **can receive** effective treatment timely and **accurately.**
- **Use differentmachinelearningmethods** to analyze **different parts of the** hospital to determine their **coverage.Hospitals** are **classified according** to their ability to **receivehigh-risk** patients.
- Machine **learningcan help** healthcare **providers** understand **theircustomers' needs,** preferences, **behaviors,patterns,** and quality in order to **improverelationships** with them.
- **Theverification system** is **basedon icasmachinelearning** techn**ology and can**

Annals of R.S.C.B., ISSN:1583-6258, Vol. 24, Issue 2, 2020, Pages. 748- 763
Received 24 October 2020; Accepted 15 December 2020

identify unknown or **abnormal** patterns in infection control data. **Association** rules are used to **generate** unexpected and interesting information from public **monitoring** and hospital **monitoring** data.

- Health **insurance companies aredeveloping** a model **that uses machine learning methods to** detect fraud and abuse in medical claims. **This** model **canbeused to** identify **inappropriate** prescriptions, irregular or **incorrect** medical claims **patterns fromdoctors,** patients, **hospitals,** etc.
- **The American** Health **Ways** system **usesmachinelearning technology to create predictive** models to **identify high-risk** patients. **The** main **purpose** of **the** system is to **manage** diabetic patients, improve their health **quality,** and **save** service **costsforpatients.Identifypatientswhoneed** more **attentionthan** other patients
- Machine **learningplays** an important role **in formulating** effective health **policies** to improve health quality **and** reduce **healthcarecosts.**



We also analyzed the GLOBOCAN 2020 database to estimatecancer incidence and mortality in different geographic regions of the world. The GLOBOCAN 2020 database was created using alarge amount of data from the Descriptive Epidemiology Group of the International Agency for Research on Cancer. (International Agency for Research on Cancer). The Cancer Registry provides incidence data. They cover the entire country's population or a sample of such populations from selected regions. The CancerRegistry also provides statistics on the survival rate of cancerpatients. Mortality data for many countries are available through the registry. These estimates are based on the latestmorbidity, mortality, and survival data providedby IARC, but thelatestdatacan beobtained directly from local sources. 0 to 14, 15 to 44, 45 to 54, 55 to 64, and ≥65 years old group. The five age groups are0.31.0.43, 0.11, 0.08, and 0.07

VI. DATA ANALYSIS OFCANCER WORLDWIDE

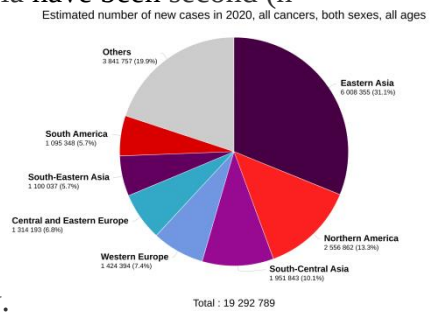
Accordingto the GLOBOCAN database, there are 19,292,789 newlyregistered cancer cases worldwide in 2020 (excluding skin cancer); of these, 10,065,305 (58.4%)aremales and 9,227,484 (48.6%)arefemales. Almost48% of new cases.In Asia, 22%are in Europe, 12.8%are in North America, 8.1%are in Latin America, and 6.9%are in Africa. The second most common location is the colon (1,065,960men and 865,630 women), followed by the stomach (719,523 men and 369,580 women).Among women, breast cancer ranksfirstwith 2,261,419 new cases per year, followed by cervicalcancer (604,127 cases) and colon cancer

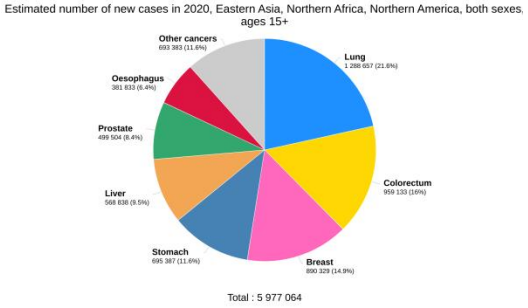
Annals of R.S.C.B., ISSN:1583-6258, Vol. 24, Issue 2, 2020, Pages. 748- 763
Received 24 October 2020; Accepted 15 December 2020

(865,630 cases). In men, the three most common cancer sites are thelungs (1,435,943 cases) and theprostate.(1,414,259 cases) and colorectal (1,065,960 cases).
The global death tollfrom cancer in 2020 is 9,958,133, including 5,528,810 males and 4,429,323 females. Lung cancer is the cause of most cancer deaths worldwide.The total number of deathsfrom lung cancer in 2002 was 1,796,144, of which 1,188,679 were men and 607,465 were women; rectal cancer and coloncancer ledto a total of 935,173 deaths, including 515,637 men and 419,536 women; liver cancer was cancer Thethird leading cause of death;a total of 830,180 deaths (577,637 men and 252,658 women)in 2020 arerelated to liver cancer. Among women, the top three causesof cancer deaths are breast cancer (684,996 deaths), lung cancer (607,465 deaths) and colon cancer (419,536 deaths), while lung cancer (1,188,679 deaths), liver (577,522), and colon Cancer and rectal cancer (515,637 people) are thethree most common cancer deaths in men.

INCIDENCE OF CANCER BY GEOGRAPHICREGIONS

Of the 21 areasindexedwithinside the GLOBOCAN 2020 database, East Asia had the biggestquantity of incident most cancers cases (all ages, all webweb sitesbesides skin) in 2020 (n = 6,008,355); NorthAmerica and South Central Asia have been second (n = 2,556,862) and third (n = 1951,843) at the list, respectively.





Thepatternofcancersitesvaried substantially from region to region.For example, the 3maximumnot unusualplacemost cancerswebweb sitesamongstpeople 15 years or older in East Asia have been lung(16.nine percent) colorecturm (12.7 percent) and stomach (eleven percent), whilethe ones in North America have been prostate (16.five percent), breast (14.7 in keeping with- cent), and lung (14.five percent) (Figures 2-3). For eachadult males and females, the prevalenceprice of most cancersacceleratednotably with age. For example, the yearly male most cancersprevalencewithinside the age organization of zeroto fourteen years became 6.forty fivein keeping with 100,000 in Western Africa,nine.07 in keeping with 100,000 in Eastern Asia,14.10 in keeping with 100,000 in Western Europe, and 15.12 in keeping with 100,000 in North America; the priceswithinside theequalareas for people whohave beensixty five years or older have been 385.44, 1461.59, 2327.87 and 2958.14 in keeping with 100,000, respectively (Table 1). North America, Australia/New Zealand, and Europe had the very bestbasicprevalenceprices in 2002, even as Northern and Western Africa had the bottomprevalenceprices (Tables 1-2). The geographic variantbecameas a

substitute substantial. For example, the age-standardized price in North American adult males (398.fourin keeping with 100,000 person-years) became4instances of the age-standardized price in North African adult males (ninety ninein keeping with 100,000 person-years).

The geographic difference in cancer incidence can be explained toa largeextent by differentsocio-economic, environmental and lifestyle factors in different parts of the world. Compared with developed countries, developing countries generally lack the resources to detect cancer cases. In developed countries, many cases of breast, prostate, colon, and cervical cancer are detected through screening tests (such as mammography,prostate-specific antigen testing, colonoscopy, and Pap smear).Chinais carrying out extensive testingwork.Usuallyrare.They also play a role, but the dominant role of genetics is only observed in a relatively small proportion of the population. It is believed that most cancers (more than90%) are causedbyacombination of genetic variation, environmental factors, and lifestyle factors [2].Exceptfor exposure to sunlight and vitamin D metabolism associatedwith cancer risk, geographic environment may have little effect on cancer risk. Themain categories of cancer risk factors include tobacco use, occupational exposure, pollution, sources of infection, and lifestyle factors.

Incidence	
1	National (or local with coverage greater than 50%) rates projected to 2020
2a	Most recent rates from a single registry applied to 2020 population
2b	Weighted/simple average of the most recent local rates applied to 2020 population
3a	Estimated from national mortality estimates by modelling, using mortality:incidence ratios derived from country-specific cancer registry data
3b	Estimated from national mortality estimates by modelling, using mortality:incidence ratios derived from cancer registry data in neighbouring countries
4	"All sites" estimates from neighbouring countries partitioned using frequency data
9	No data: the rates are those of neighbouring countries or registries in the same area
Mortality	
1	National rates projected to 2020
2a	Most recent rates from one source applied to 2020 population
2b	Weighted/simple average of the most recent local rates applied to 2020 population
3	Estimated from national incidence estimates by modelling, using incidence:mortality ratios derived from cancer registry data in neighbouring countries
9	No data: the rates are those of neighbouring countries in the same area

Tobacco use

Over the years, people have accumulated knowledge about the role of smoking in the etiology of cancer [3]. In 2004, the International Agency for Research on Cancer published a monograph on tobacco use and cancer, which concluded that tobacco more or lesscausescancer.In 15 different locations, including lungs, urinary tract, upper respiratory tract, pancreas, stomach and liver. Although smoking rates in many developed countries

Annals of R.S.C.B., ISSN:1583-6258, Vol. 24, Issue 2, 2020, Pages. 748- 763
Received 24 October 2020; Accepted 15 December 2020

have declined,smokingrates in developing countries are on the rise [4]. Currently,about 5 million people diefrom tobacco use each year; estimates based on current trends indicate that this number will increase to 10 million by 2030,and70% of deaths will occur in developing countries. [5] It is important to takemeasures such as increasing taxes on tobaccoproducts,disseminating information about the health risks of smoking, restricting smoking in public places and workplaces, extensive advertising and promotion bans, and expanding the useof smoking cessation methods to reduce the incidence of cancer . And other tobacco-related diseases [5].

Table 3. Age-specific mortality rates of all cancers (except skin) among males in 2020*.

	0-14	15-44	45-54	55-64	65+	All ages	ASR**
Eastern Africa	10.66	59.17	204.22	380.4	736.85	73.9	133.2
Middle Africa	7.11	49.73	198.74	375.94	646.45	65.6	120.8
Northern Africa	7.28	19.81	111.83	293.42	522.28	53.9	83.1
Southern Africa	6.13	38.82	240.95	584.94	951.5	94.5	158.5
Western Africa	4.14	28.02	152.52	244.21	340.43	41.3	73.5
Caribbean	5.98	16.74	120.01	343.39	1229.53	123.8	135.8
Central America	6.06	14.61	75.72	230.21	859.4	63	95.1
South America	6.35	18.4	131.69	369.27	1112.8	101.1	131.8
North American	2.65	16.3	128.69	417.08	1394.91	210.2	153
Eastern Asia	5.05	31.07	229.81	471.03	1198.95	167	161.8
Southeastern Asia	6.87	19.89	147.18	339.23	692.48	73	102.5
South Central Asia	4.3	15.79	113.69	283.14	495.86	55.2	78
Western Asia	7.27	18.86	122.48	379.68	778.81	74.4	108.7
Eastern Europe	5.53	28.64	256.3	750.41	1356.35	253.5	197.2
Northern Europe	3.33	14.52	131.68	413.01	1517.66	269.9	161
Southern Europe	3.47	20.38	175.51	495.03	1447.52	294.4	170.1
Western Europe	3.09	19.58	177.63	483.88	1518.44	294.6	173.9
Australia/New Zealand	3.7	15.37	110.58	378.27	1412.69	213.9	149.1
Melanesia	5.7	24.69	152.6	358.93	667.37	57.3	104.6
Micronesia	7.97	23.48	128.46	323.52	884.31	80.4	114.5
Polynesia	3.45	20.46	203.46	529.58	738.05	86.1	126.3

Table 4. Age-specific mortality rates of all cancers (except skin) among females in 2020*.

	0-14	15-44	45-54	55-64	65+	All ages	ASR**
Eastern Africa	7.15	60.04	256.42	408.34	482.72	75	122.7
Middle Africa	5.04	52.21	213.85	313.89	375.8	61.5	99
Northern Africa	5.27	29.24	135.05	208.92	276.45	50.2	65.1
Southern Africa	4.67	40.94	219.9	364.99	483.42	81.1	106.3
Western Africa	3.3	41.77	198.17	264.1	253.65	50.2	79.7
Caribbean	5.01	26.66	140.65	269.81	690.16	100.3	98.4
Central America	5.22	20.72	133.72	253.04	630.68	67.9	89.6
South America	5.19	23.87	148.37	285.8	731.22	93.4	102.2
North American	2.32	19.4	124.07	321.35	910.18	185.8	112.1
Eastern Asia	3.74	20.79	134.74	250.45	590.09	100.8	86.3
Southeastern Asia	5.47	24.98	156.44	244.17	385.43	62.5	76.2
South Central Asia	2.79	22.32	151.7	265.77	307.54	55.2	69.9
Western Asia	5.82	21.36	123.79	238.9	432.93	57.7	74
Eastern Europe	4.56	28.97	158.84	326.95	633.93	175.7	101.9
Northern Europe	2.6	20.44	138.34	327.26	958.56	236.9	118.1
Southern Europe	3.04	19.93	119.1	244.68	714.06	189.03	92.2
Western Europe	2.21	20.7	127.62	274.95	864.06	224.9	106.1
Australia/New Zealand	2.67	19.18	116.09	283.8	841.32	167.1	103.4
Melanesia	3.84	37.06	233.53	400.91	424.78	66.5	104.6
Micronesia	4.12	29.33	175.51	261.52	492.51	66.8	88.6
Polynesia	4.81	48.83	268.4	247.04	368.27	79.2	97.6

*Rates are per 100,000 person-years. **Age-standardized rates using the world standard.

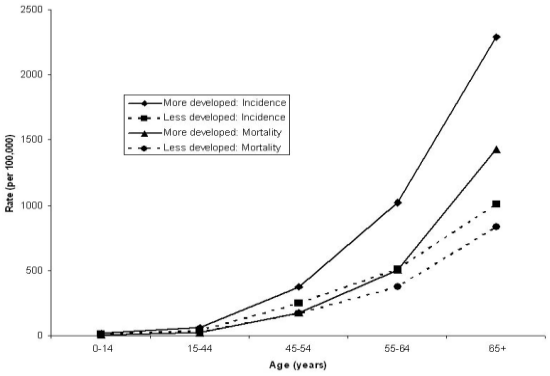
Occupational exposures

For a long time, occupational exposure has beenassociatedwith cancer risk. A recent publication lists 28 important occupational human carcinogens, ranging from ionizing radiation, asbestos, silica, wood chips and arsenic to benzene [6]. In general,industrialized countries have experienced the process of industrialization earlier than developing countries, and people living in industrialized countries are oftenmoresusceptibleto theinfluence of various occupations. Concerned about the lack of resources to monitor occupational exposure and establish or enforce occupational standards, this is usually an ongoing process. For example, the current professional standard for benzene, one of the

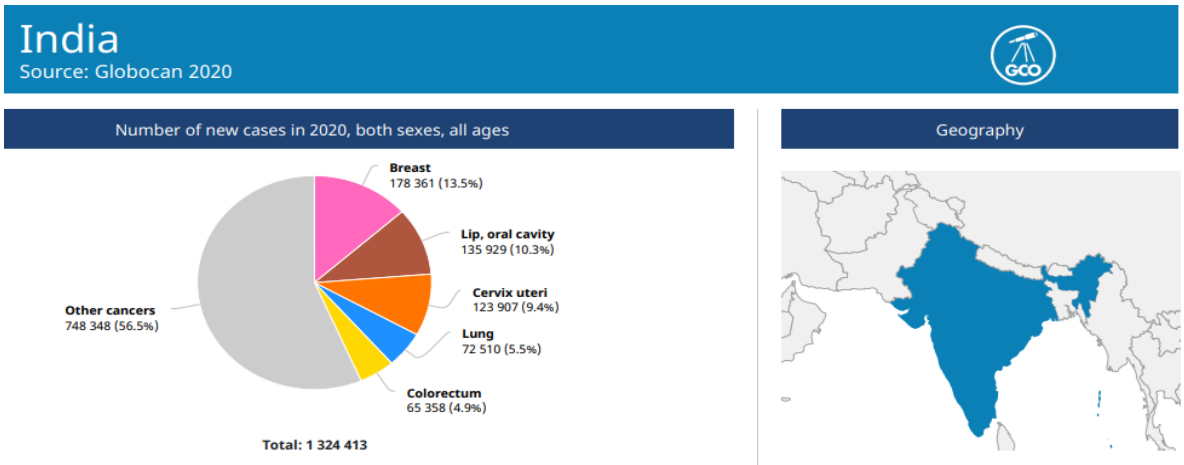
Annals of R.S.C.B., ISSN:1583-6258, Vol. 24, Issue 2, 2020, Pages. 748- 763
Received 24 October 2020; Accepted 15 December 2020

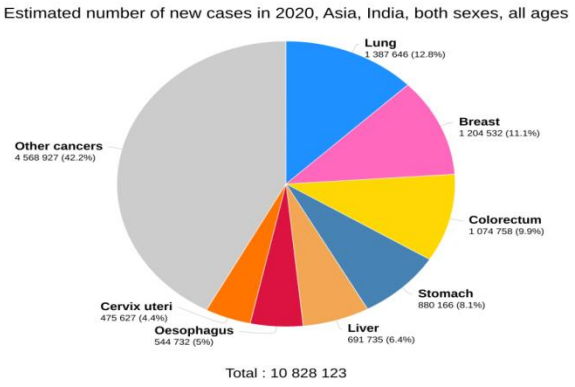
most widely used industrial chemicals and known carcinogens in the United States, is one part per million (ppm) or 3.26 mg/m3.

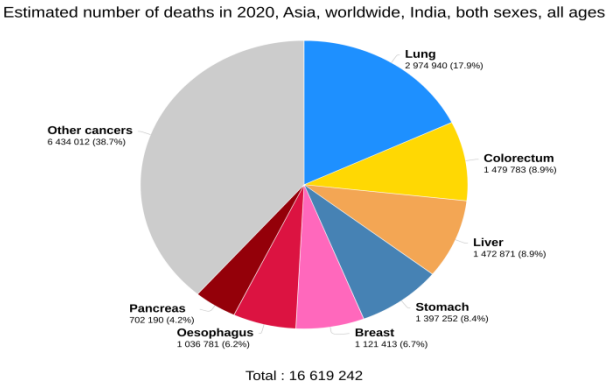
Figure 4.Male age-specific incidence and mortality in more or less developed countries, 2020.

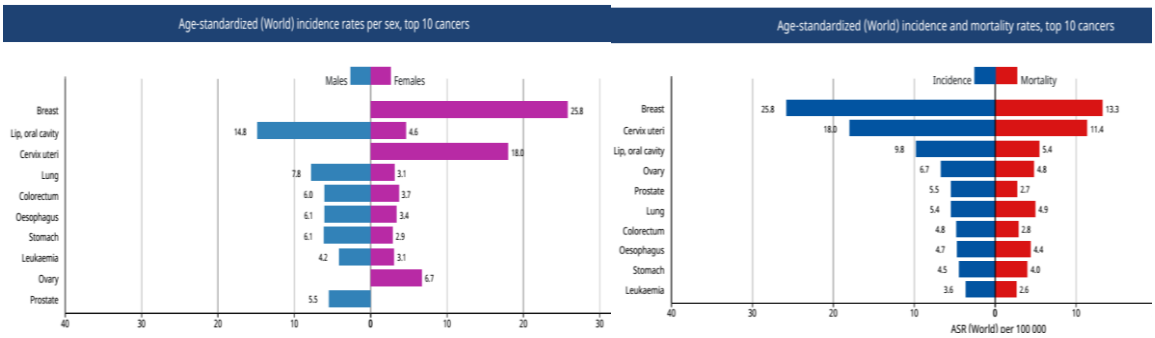


Results





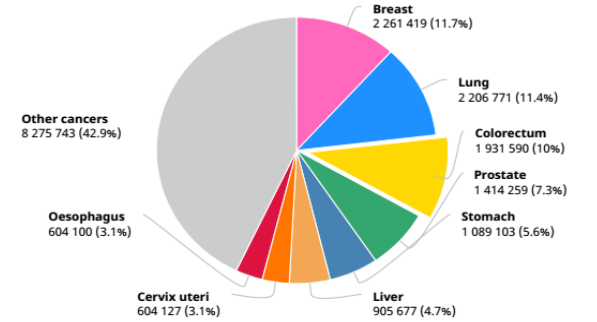




Annals of R.S.C.B., ISSN:1583-6258, Vol. 24, Issue 2, 2020, Pages. 748- 763
Received 24 October 2020; Accepted 15 December 2020

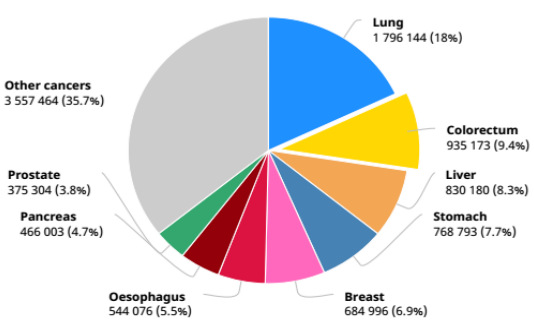
Colorectal Cancer Results

Number of new cases in 2020, both sexes, all ages



Total: 19 292 789 cases

Number of deaths in 2020, both sexes, all ages

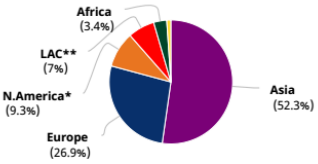


Total: 9 958 133 deaths

Cancer incidence and mortality statistics worldwide and by region

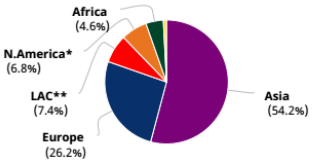
	Incidence						Mortality					
	Both sexes		Males		Females		Both sexes		Males		Females	
	New cases	Cum. risk 0-74 (%)	New cases	Cum. risk 0-74 (%)	New cases	Cum. risk 0-74 (%)	Deaths	Cum. risk 0-74 (%)	Deaths	Cum. risk 0-74 (%)	Deaths	Cum. risk 0-74 (%)
Eastern Africa	18 306	0.92	8 888	1.00	9 418	0.85	13 236	0.67	6 365	0.74	6 871	0.62
Middle Africa	5 767	0.80	3 045	0.90	2 722	0.72	4 228	0.59	2 222	0.67	2 006	0.53
Northern Africa	20 858	1.10	10 662	1.17	10 196	1.03	11 530	0.56	5 900	0.60	5 630	0.52
Southern Africa	7 684	1.57	3 919	1.93	3 765	1.30	3 943	0.79	2 052	1.00	1 891	0.64
Western Africa	13 583	0.74	7 546	0.88	6 037	0.63	9 938	0.56	5 507	0.66	4 431	0.47
Caribbean	11 454	2.04	5 327	2.13	6 127	1.97	6 983	1.08	3 307	1.18	3 676	0.98
Central America	19 535	1.19	10 181	1.37	9 354	1.03	10 439	0.61	5 494	0.72	4 945	0.51
South America	103 954	2.09	51 710	2.35	52 244	1.86	52 013	0.95	26 175	1.11	25 838	0.82
Northern America	180 575	2.98	95 138	3.38	85 437	2.61	63 987	0.87	34 105	1.05	29 882	0.70
Eastern Asia	757 849	3.00	431 501	3.64	326 348	2.37	368 072	1.23	208 090	1.57	159 982	0.91
South-Eastern Asia	106 995	1.70	60 505	2.11	46 490	1.33	57 064	0.82	32 205	1.06	24 859	0.62
South-Central Asia	102 987	0.63	61 252	0.76	41 735	0.50	59 206	0.36	35 848	0.44	23 358	0.27
Western Asia	41 569	1.97	23 496	2.35	18 073	1.63	22 107	0.97	12 382	1.18	9 725	0.79
Central and Eastern Europe	172 950	3.63	89 189	4.75	83 761	2.85	93 384	1.72	48 378	2.38	45 006	1.26
Western Europe	141 644	3.27	77 052	3.95	64 592	2.65	62 266	1.06	34 113	1.37	28 153	0.77
Southern Europe	123 588	3.76	71 009	4.81	52 579	2.81	55 406	1.22	31 583	1.60	23 823	0.88
Northern Europe	81 638	3.89	44 464	4.52	37 174	3.31	33 768	1.16	17 811	1.38	15 957	0.96
Australia and New Zealand	19 644	3.70	10 491	4.24	9 153	3.18	7 038	0.87	3 755	1.13	3 283	0.63
Melanesia	804	1.32	466	1.66	338	1.01	452	0.75	279	0.99	173	0.53
Polynesia	113	1.90	68	2.41	45	1.40	60	1.14	37	1.36	23	0.93
Micronesia	93	2.07	51	2.55	42	1.63	53	1.21	29	1.64	24	0.84
Low HDI	37 923	0.84	19 616	0.94	18 307	0.76	27 443	0.62	14 173	0.69	13 270	0.55
Medium HDI	129 206	0.70	74 698	0.85	54 508	0.56	74 912	0.40	43 769	0.49	31 143	0.31
High HDI	812 972	2.39	455 544	2.86	357 428	1.94	421 087	1.10	237 016	1.38	184 071	0.85
Very high HDI	950 563	3.42	515 616	4.15	434 947	2.77	411 289	1.15	220 443	1.47	190 846	0.87
World	1 931 590	2.25	1 065 960	2.71	865 630	1.83	935 173	0.94	515 637	1.18	419 536	0.73

Incidence, both sexes



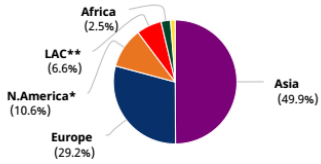
Population	Number
Asia	1 009 400
Europe	519 820
*Northern America	180 575
**Latin America and the Caribbean	134 943
Africa	66 198
Oceania	20 654
Total	1 931 590

Mortality, both sexes



Population	Number
Asia	506 449
Europe	244 824
**Latin America and the Caribbean	69 435
*Northern America	63 987
Africa	42 875
Oceania	7 603
Total	935 173

5-year prevalence, both sexes



Population	Number
Asia	2 622 900
Europe	1 536 168
*Northern America	554 680
**Latin America and the Caribbean	345 715
Africa	130 338
Oceania	63 534
Total	5 253 335

VII. CONCLUSION

The prediction of Cancer Dieses is very important aspect in Health care sector. This paper, describes the various Machine Learning Techniques,whichare used to predict CancerDiseases. It concludes that there is no single Machine Learning Technique which gives consistent results for all type of Cancer Diseases. The performance of the Machine Learning Techniques depends on the type of data setthat is

used in medical diagnosis. The main idea of this survey is using different Machine Learning Techniques on Cancer Diseases yields different results. These comparing results give best algorithm for futurework.

VIII. FUTURERESEARCH

Cancer is primarily a disease of the elderly;if all other factors remain the same, demographic changes (population growth and an increase in the proportion of elderly people in the world's population) will lead to an increase in the global cancer incidence. A method of predicting the incidence of cancer in the future. Thisemphasizes the importance ofimproving our understanding of cancer risk factors, formulating and implementing practical prevention strategies, and developing better and more effective treatment options. And less developed countries In less developedcountries, the proportion of elderly people is low. As the incidence of cancerincreases with age, the population base of underdeveloped countries is larger, andthere maybe more room for an aging population.In less developed countries, it will be even lower. The National Academy of Sciences InstituteofMedicine recently published "Cancer Control Opportunities in Developing Countries"[22] that cancer is a major disease burden in low- and middle-income countries, and this burden is becoming more andmore serious, not just Forthese countries, because these countries havea largerpopulation,there are more cases, but there are also more aggressive cancers and a lower cure rate. The report also pointedout thatthe causes and consequencesof cancervarygreatly between developed and underdeveloped countries. For example, in developing countries, one in four cancer cases are related to the source of infection,while in developedcountries, this proportionis less than one in ten.These differencesindicate that moreand more countries and less developedcountrieshave different cancer prevention and control strategies in the allocationof resources in thisfield.Cancer etiology, prevention, treatment and health policy research to reduce the global burden of cancer.

-VII. REFERENCES

[1] Satyam Shukla, DharmendraLalGuptaandBakshiRohit Prasad, " Comparative Study of Recent Trends on Cancer Disease Prediction using Data Mining Techniques",International Journal of Database Theory and Application Vol.9, No.9 (2016), pp.107-118.

[2] GCO. https://gco.iarc.fr/today/data/factsheets/cancers/10_8_9-Colorectum-fact-sheet.pdf

[3] K. Balachandran and R. Anitha, "Ensemble based optimal classification model for pre-diagnosis of lung cancer", 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT), IEEE, (2013).

[4] K. Shiny, "Implementation of Data Mining Algorithm to Analysis Breast Cancer", International Journal for Innovative Research in Science and Technology, vol. 1, no. 9, (2015),pp.207-212.

[5] S. S. Shrivastava, V. K. Choubey and A. Sant, "Classification Based Pattern Analysis on the Medical Data in Health Care Environment", International Journal of Scientific Research in Science, Engineering and Technology, vol. 2, no. 1, (2016).

[6] R. Vidhu and S. Kiruthika, "A New Feature Selection Method for Oral Cancer Using Data Mining Techniques", International Journal of Advanced Research in Computer and Communication Engineering, vol. 5, no. 1, (2016).

[7] D. Tomar and S. Agarwal, "A survey on Data Mining approaches for Healthcare", International Journal of Bio-Science and Bio-Technology, vol. 5, no. 5, (2013), pp. 241-266.

[8] H. Li, G. Hong and Z. Guo, "Reversal DNA methylation patterns for cancer diagnosis", 2014 8th International Conference on Systems Biology (ISB), IEEE, (2014).

[9] R. Chau, "Determining the familial risk distribution of colorectal cancer: a data mining approach", Familial cancer, (2015), pp. 1-11.
[10] N. Rathore, D. Tomar and S. Agarwal, "Predicting the survivability of breast cancer patients using ensemble approach", 2014 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT), IEEE, (2014).
[11] Berman AT, James SS, Rengan R. Structure, mechanism, and evolution of the mRNA capping apparatus. Cancers (Basel). 2015;7(3):1178–90.
[12] D. Tomar and S. Agarwal, "Hybrid feature selection based weighted least squares twin support vector machine approach for diagnosing breast cancer, hepatitis, and diabetes", Advances in Artificial Neural Systems, vol. 2015, no. 1.
[13] B. R. Prasad and S. Agarwal, "Modeling risk prediction of diabetes-A preventive measure", 9th International Conference on Industrial and Information Systems (ICIIS), Gwalior, IEEE, (2014).pp.1-6.
[14] D. Tomar, B. R. Prasad and S. Agarwal, "An efficient Parkinson disease diagnosis system based on Least Squares Twin Support Vector Machine and Particle Swarm Optimization", 9th International Conference on Industrial and Information Systems (ICIIS), Gwalior, IEEE, (2014), pp. 1-6.
[15] A. K. Yadav, D. Tomar and S. Agarwal, "Clustering of lung cancer data using Foggy K-means", 2013 International Conference on Recent Trends in Information Technology (ICRTIT), IEEE, (2013).
[16] Chen H, Zhang H, Zhang Z, Cao Y, Tang W. Network-based inference methods for drug repositioning. Comput Math Methods Med. 2015;2015:130620.
[17] Lee HS, Bae T, Lee JH, Kim DG, Oh YS, Jang Y, et al. Rational drug repositioning guided by an integrated pharmacological network of protein, disease and drug. BMC Syst Biol. 2012;6:80.
[18] HuangCH,WuMY,Chang PM, Huang CY, Ng KL.Insilicoidentificationofpotential targetsanddrugsfornon-smallcelllung cancer. IETSystBiol.2014;8(2):56–66.
[19] Huang CH, Chang PM, Lin YJ, Wang CH, Huang CY, Ng KL. Drug repositioning discovery for early- and late-stage non-small-cell lung cancer. Biomed Res Int. 2014;2014:193817.
[20] Huang CH, Peng HS, Ng KL. Prediction of cancer proteins by integrating protein interaction, domain frequency, and domain interaction data using machine learning algorithms. Biomed Res Int. 2015;2015:312047
[21] Stachnik A, Yuen T, Iqbal J, Sgobba M, Gupta Y, Lu P, et al. Repurposing of bisphosphonates for the prevention and therapy of nonsmall cell lung and breast cancer. ProcNatlAcadSci U S A. 2014;111(50):17995–8000.
[22] Kamangar F, Dores GM, Anderson WF. Patterns of cancer incidence, mortality, and prevalence across five continents: defining priorities to reduce cancer disparities in different geographic regions of the world. J ClinOncol. 2006;24(14):2137-50.
[23] Institute of Medicine (U.S.) Committee on Cancer Control in Low- and Middle-Income Countries; Sloan FA, Gelband H, editors. Cancer control opportunities in low- and middle-income countries. Washington, DC: National Academies Press; 2007
[24] Parkin DM, Bray F, Ferlay J, Pisani P. Global cancer statistics, 2002. CA Cancer J Clin. 2005;55(2):74-108.
[25] Parkin DM. International variation. Oncogene. 2004;23(38):6329-40.
[26] Doll R, Peto R, Boreham J, Sutherland I. Mortality from cancer in relation to smoking: 50 years observations on British doctors. Br J Cancer. 2005;92(3):426-9.
[27] World Health Organization. IARC. Tobacco smoke and involuntary smoking : Views and expert opinions of an IARC working group on the evaluation of carcinogenic risks to humans, which met in Lyon; 2002 11-18 June. Lyon: IARC. 2004.

[28] Jha P, Chaloupka FJ, Corrao M, Jacob B. Reducing the burden of smoking world-wide: effectiveness of interventions and their coverage. Drug Alcohol Rev. 2006;25(6):597-609.
[29] Siemiatycki J, Richardson L, Straif K, et al. Listing occupational carcinogens. Environ Health Perspect. 2004;112(15):1447-59.
[30] Liang Y, Wong O, Yang L, Li T, Su Z. The development and regulation of occupational exposure limits in China. RegulToxicolPharmacol. 2006;46(2):107-13.
[31] Lan Q, Zhang L, Li G, et al. Hematotoxicity in workers exposed to low levels of benzene. Science. 2004;306(5702):1774-6. *Ma and Yu: Global Burden of Cancer 93*
[32] Parkin DM. The global health burden of infection-associated cancers in the year 2002. Int J Cancer. 2006;118(12):3030-44.
[33] Uemura N, Okamoto S, Yamamoto S, et al. Helicobacter pylori infection and the development of gastric cancer. N Engl J Med. 2001;345(11):784-9.
[34] Hoofnagle JH, Doo E, Liang TJ, Fleischer R, Lok AS. Management of hepatitis B: Summary of a clinical research workshop. Hepatology. 2007;45(4):1056-75.
[35] MagalhaesQueiroz DM, Luzza F. Epidemiology of Helicobacter pylori infection. Helicobacter. 2006;11Suppl 1:1-5.
[36] Stuver SO, Boschi-Pinto C, Trichopoulos D. Infection with hepatitis B and C viruses, social class and cancer. IARC Sci Publ. 1997;138:319-24.
[37] Ogden CL, Carroll MD, Curtin LR, McDowell MA, Tabak CJ, Flegal KM. Prevalence of overweight and obesity in the United States, 1999-2004. JAMA. 2006;295(13):1549-55.
[38] Calle EE, Kaaks R. Overweight, obesity and cancer: epidemiological evidence and proposed mechanisms. Nat Rev Cancer. 2004;4(8):579-91.
[39] Kahn SE, Hull RL, Utzschneider KM. Mechanisms linking obesity to insulin resistance and type 2 diabetes. Nature. 2006;444(7121):840-6.
[40] Bray GA, Bellanger T. Epidemiology, trends, and morbidities of obesity and the metabolic syndrome. Endocrine. 2006;29(1):109-17