Ilkogretim Online - Elementary Education Online, 2020; Vol 19 (Issue 4): pp. 6915-6925
http://ilkogretim-online.org
doi: 10.17051/ilkonline.2020.04.765097

Predicting The Risk Of Cancer By Diagnosing Tumor

Mr.Parthiban M Assistant Professor, Department of Computer Science and Engineering, Sri Sai Ram Institute of Technology, Chennai

Ms.Rajalakshmi G Student, Department of Computer Science and Engineering, Sri Sai Ram Institute of Technology, Chennai

Abstract-: When cells grow and divide more than they should or do not die when they should, an abnormal mass of tissue arises. Tumors can be benign (noncancerous) or malignant (cancerous). Benign can develop to be quite large, but they would not spread to healthy tissue or other sections of the body. Malignant tumors can invade or disseminate into surrounding structures.

They can also spread through the blood and lymph systems to other regions of the body. Also termed as neoplasm.. But research has shown that specific risk factors may change the person's chances of producing human. Cancer be prevented by the diagnosis of tumor. As tumor can either be benign or malignant. By proper diagnosis cancer development and probability of future occurrence of tumor can bedetected.

Keywords: tumor, cancer, Markov chain, artificial neural network, Naïve Bayes.

1. INTRODUCTION

The basic units that make up the human body are cells. Cells divide and expand to produce new cells as the body requires them. Cells normally die when they get too old or damaged. Then, in their place, new cells appear.

When genetic alterations disrupt this normal cycle, cancer develops. Cells begin to proliferate at an uncontrollable rate. These cells may clump together to form a tumor. Tumors can be malignant or noncancerous. A malignant tumor is one that has the potential to grow and spread to other regions of the body. The term "benign tumor" refers to a tumor that can develop but not spread. A tumor is a tissue growth or lump that resembles swelling. Although not all tumours are cancerous, seeing a doctor if one emerges is a good idea. Tumors can range in size from a little nodule to a massive mass, and they can appear practically anywhere on the body, depending on thetype.

As a result, the problem of monitoring and predicting cancer-disease development has emergedas a major challenge and research focus.C omputer science and information technologies have advanced dramatically over the last two decades, and they are now playing an important role in cancer research. Because of their superior performance in simulation and modelling, data mining and machine learning approaches are increasingly beingused.

For example, 2018 et al Heidari proposed a machine learning-based method. Their work to centralise and identify breast cancer tissue was effective in classifying incoming medical images into malignant, benign, and healthy patients. Other successful ANN-based model implementations can be found in Siddiquiet survey. .'s(2020).

The majority of traditional ANN applications, on the other hand, consider network inputs directly from the original data, with less work done in terms of input amendment or augment. The standard network training process, on the other hand, is usually time consuming, especially when there are a large number of inputs. Furthermore, in some real-world scenarios, the standard ANN's generalisation performance is far from satisfactory. Chaoyu Yang's review looks in to the identifying subsequent disease development using a hybrid algorithm of ANN, Naive Bayes, andMarkov chain (2020). To address the issue of predicting patients' disease development, we try to predict the probability of reoccurrence of tumour by identifying its type and medical conditions using an existing algorithm based on the idea of Artificial Neural Network, Navie Bayes, and Markov chain.

The remaining portion of the paperwork is outlined. Section 2 summarises the existing work and the results obtained. Section 2.1, 2.2, 2.3 provides a review of the literature on the topics analysedin the domain of cancer risk analysis ANN, Nave Bayes, and Markov chain. Section 3 provides background information on the research, such as a description of the target dataset used in this study and the Markov chain. Section 4 provides the proposed approach and the work flow of the study. Section 5 then discusses the subsequent disease development and Section 6 describes t the existing ANN model, Section 7 describes the overview and experimental outcomes of the project and Section 8 concludes thestudy.

2. EXISTING WORK

Cancer-related studies, such as patient status monitoring, medical resource allocation, and survivability prediction, to name a few, have attracted a lot of attention (Loud and Murphy, 2017). Heidari et al. (2018) proposed amachine learning-based model to identify mammographic image features for short- term breast cancer prediction in their work. The results also demonstrated a significant improvement in their work when compared to standard methods such as Linear Regression and Decision Tree. In addition, a comparison of the Nave Bayes and K-Nearest Neighbor. In the medical domain, we have seen a large number of ANN-based applications. For example, Fakooret al. (2013) developed a hybrid method for cancer detection that combined ANN with the Support Vector Machine and was tested on several gene-expression datasets. Amrane et al.(2018) provided (KNN) algorithms for the categorization of breast cancer. An investigation combining the Bayesian Network and Markov Chain models to modify the Artificial Neural Network's input. The proposed algorithm (in Chand Yoag 2020) is then applied to one of the world's largest cancer-related datasets, and a comparison with state-of- the-art approaches ismade.

2.1 Prediction of Cancerdevelopment

Cancer risk assessment is critical for healthcare providers and medical researchers. Several research studies have attempted to provide a diverse range of cancer risk management and/or prediction strategies. The ultimate goal is to provide precautionary measures for people who are at risk, as well as to monitor disease progression (or survivability prediction).

Hart et al. (2018) used a multi- parameterized neural network for lung cancer risk prediction, which was based on putative risk factors as well as clinical data and also demographic data

Despite the high level of interest in cancer risk and survivability analysis, little research has been conducted on the relationship between patients' previous and current diagnoses. This research question is critical because it aids in the prediction of patients' future disease development. Gaining a thorough understanding of the potential risk for subsequent diseases also aids in improving healthcare quality and treatment services. In (Chand Yoag 2020), a probabilistic model that incorporates the techniques of the Artificial Neural Network, Naïve Bayes, and the Markov chain model is proposed. This study aims to investigate the early stage of tumour by identifying its type and the likelihood of tumour cell development to cancer by analysing previous medical reports and family healthhistory.

2.2 Artificial neuralnetwork

The Artificial Neural Network (ANN) is a popular data-mining algorithm that can respond to complex inputs and generate desired outputs. Because of its satisfactory performance and high accuracy, ANN has found widespread application in a variety of fields, including pattern recognition prediction, statistical simulation, and so on. The artificial neuron is the most fundamental computing unit in ANN. These neurons are built in the same way that biological neurons in the human brain are. In general, input signals aresent.

Assume the i-th neuron's input signal is a vector of xi, the connection strength to the output isthe weight wj, and its bias input isrepresented byb.

Given the activation function f (.), the output of the i-th neuron is as follows:

$$y = f(x^T iwi + b)$$

Furthermore, after determining the activation function and network structure, a training process is required to update the internal network weights in order to minimise the error between the actual network and the desired output, as done in the Chand Yoag 2020. Back Propagation, Resilient Propagation, and other common learningalgorithms.

2.3. Naïve Bayes and Markov Chain

From [1] Bayesian theory provides a computational framework for estimating conditional probability, which has been shown to be effective in a variety of applications. Assume we have one training sample x and n different class labels ci . The posterior probability (for x) of belonging to the i-th class [or prob(ci|x)] can then be expressed as follows: where prob(ci) denotes the class prior probability, prob(x) the prior probability of x, and prob(x|ci) the posterior probability of x given the ci classcondition.

The Markov chain model, on the other hand, is required to compute the transition probability from one state to another. The first order Markov chain, in particular, operates on the assumption that future states for one particular element (or event) depend only on the current state and not on previous states. In other words, consider xi I = 1, 2,,m) to be a sequence of random variables. The probability of transitioning to the next state (or xm+1) is then estimatedas:

The Markov chain model is useful for factoring the sequential characteristics of events.

Summary

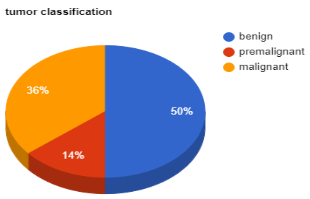
In this section, we will review some existing research on data mining techniques in the medicaldomain.

In addition, we get around a fundamental discussion of three popular methods: the Artificial Neural Network, Nave Bayes, and Markov chain model. Based on these three methods, we will employ an existing novel prediction algorithm to monitor and predict patients' disease progression by diagnosing the tumor, as discussed in the following sections.

3. STUDY BACKGROUND

Pondering for a dataset, we discerned some medical reports .This incidence database is made up of de- identified patient data from various types of cancer diseases. In addition, there are 137 features for each patient record. These features cover both demographic and clinical data. Gender, ethnicity, year of birth, month and year of diagnosis, age, and marital status of patients at diagnosis are examples of demographic information. Clinical information includes the primary site of the tumour, a tumour marker, the size of the tumor, the types of treatment received, behaviour patterns codes, laterality, andhistology.

Cancers are developed from tumors which can be of two major types. That is benign(non cancerous) and malignant(cancerous) . The earlier stage of the malignant tumor is premalignant which can beprevented from further development if diagnosedproperly.



4. PROPOSEDAPPROACH

In this section, we propose a method for predicting the probability of cancer development with different parameters by analyzing the tumor as well as the patients' previous clinical details by incorporating three different methods, including the Bayesian and Markov models, as well as the artificial neural network, which is a existing novel predictionalgorithm
. The output from the two probabilistic models will be cast into the analysis after they have been diagnosed with cancer.

4.1 DataPreprocessing

Data preprocessing is a data mining technique used to convert raw data into a usable and efficient format.

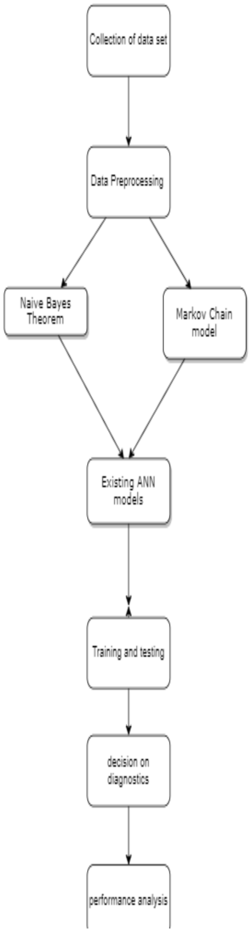
Data Preprocessing StepsIncluded:

1. Data Cleaning: There may be many irrelevant and missing parts in the data. Data cleaning is performed in order to handle this section. It entails dealing with missing data, noisy data, and soon.

2. Data transformation: The process of altering the format, structure, or values of data.

3. Data Reduction: Data reduction is a process that takes the original data and reduces it to a much smaller size. While reducing data, data reduction techniques ensure dataintegrity.

Among the 132 features, 17 independent features were chosen that may have an impact on cancer prediction tasks, such as gender, race, status, age, primary site, tumour size, colour, texture, past alignments, and so on.

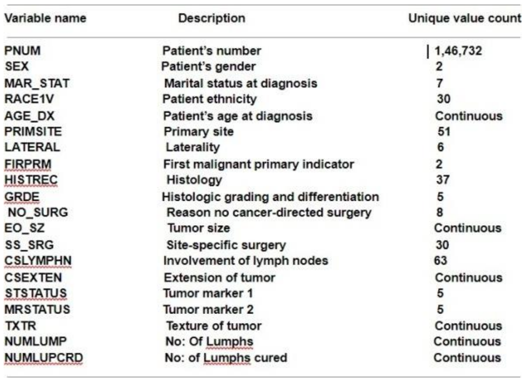


Working with the collected data, it is determined that itis impractical to backtrack those new features from previous records. In this study, patients' records with missing values will be removed for brevity's sake. That is, only completed data samples will be taken into account. Following that, we discover that selected attributes can be classified as discrete or continuous. When compared to continuous attributes, discrete attributes are easier to process. For continuous data, the minimum-maximum normalization is used, which limits the values from continuous features to the range [0, 1]. Let vpjbe the value from the p-thsample and the j-th continuous feature, and min(vj) and max(vj) be the minimal and maximal values of this j-th feature from all samples, respectively. As a result, the normalized value vpj will be estimatedas

$$\hat{v}_j^p = \frac{v_j^p - min(\boldsymbol{v}_j)}{max(\boldsymbol{v}_j) - min(\boldsymbol{v}_j)}.$$

From [1].

The attributes used are



Variable name	Description	Unique value count
PNUM	Patient's number	1,46,732
SEX	Patient's gender	2
MAR_STAT	Marital status at diagnosis	7
RACE1V	Patient ethnicity	30
AGE_DX	Patient's age at diagnosis	Continuous
PRIMSITE	Primary site	51
LATERAL	Laterality	6
FIRPRM	First malignant primary indicator	2
HISTREC	Histology	37
GRDE	Histologic grading and differentiation	5
NO_SURG	Reason no cancer-directed surgery	8
EO_SZ	Tumor size	Continuous
SS_SRG	Site-specific surgery	30
CSLYMPHN	Involvement of lymph nodes	63
CSEXTEN	Extension of tumor	Continuous
STSTATUS	Tumor marker 1	5
MRSTATUS	Tumor marker 2	5
TXTR	Texture of tumor	Continuous
NUMLUMP	No: Of Lumphs	Continuous
NUMLUPCRD	No: of Lumphs cured	Continuous

And other data used are to diagnose the development are Id,diagnosi,radius_mean,texture_mean,perimeter_mean,area_mean,smoothness_mean ,compactness_mean,concavity_mean,concavepoints_mean,symmetry_mean,fractal_dimension_mean, radius_se, texture_s, perimeter_se,area_se, smoothness_se, compactness_se ,concavity_se, concave points_se,symmmetr,fractal_dimension_se ,radius_worst,texture_worstperimeter_worst,area_worst,smoothness_worst,compactness_worst,concavity_worst, concave points_wors, symmetry_worst, fractal_dimension_worst.

5. SUBSEQUENT DISEASE- DEVELOPMENT ESTIMATION

The disease's subsequent development has already been studied in [1].

$$P(D^{p}_{i+1}|D^{p}_{i}) = \frac{N(D^{p}_{i+1}, D^{p}_{i})}{N(D^{p}_{i})},$$

N(Dp^i+1 , Dp^i) is the number of patients with a diseaseDi+1.

Markov and Nave Bayes models are used, and C(D pi, vpi) is the equality used to ensure that all probabilities sum to 1.

6. ANNModel

Overall, we investigate the existing Nave Bayes and Markov chain models to estimate the possibility of disease development. We then consider this probability result, among other original features, as an additional input for training a network. Finally, to mitigate the effects of the large number of input features, a sparse training strategy is used to maximise the network structure while simultaneously limiting the training error. Algorithm 1 sums up the proposed method for investigating cancer risk analysis to the thatend.

Algorithm 1: Existing cancer-risk prediction algorithm based on an improved

probabilistic neural network.

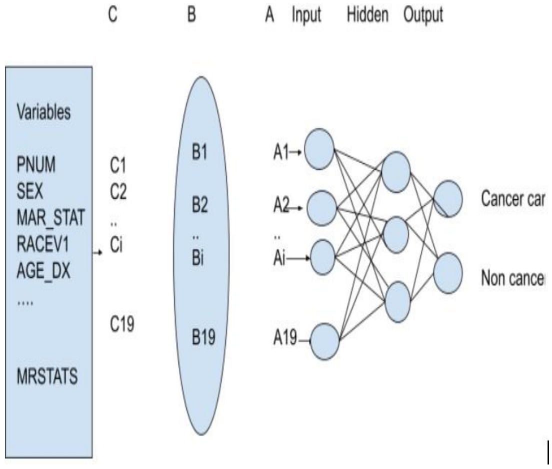
Stage 1: Data preprocessing, which includes feature selection, the removal of missing records, and data normalisation.

Stage 2: Determine the probability.

Stage 3: Use the probability result and the original input features to train the network:

Stage 3.1: Assign weights to the hidden state at random; Stage 3.2: Solve the optimization problem to obtain a spare weight matrix for the hidden-output layer.

The trained neural network should be output.

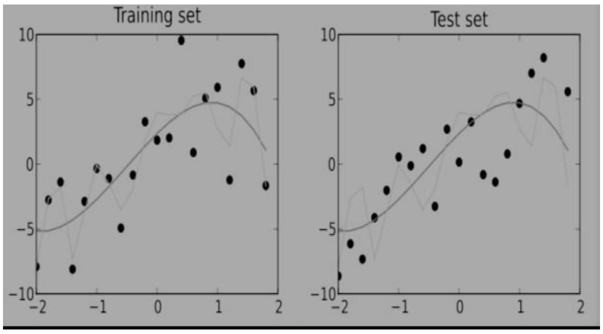


7. OVERVIEW

EXPERIMENTAL OUTCOMES

This has been described in [1] the experimental results obtained by applying the parameters to the existing algorithm to investigate the disease development of a patient. W we present the experimental setup and evaluation metrics, and we discuss the probabilities based on their historical information and individual profiles, while the performance is evaluated using the existing algorithm, which is efficient of producing thepossible

outcomes expected being trained with the available dataset. The training and the tested data results has been articulated.



8. CONCLUSION

Understanding patients' cancer risks using their historical medical data is a major focus of healthcare management. There are still many challenges to overcome, such as high dimensionality and heterogeneous data structure. In this study, a novel algorithm based on an improved probabilistic neural network is tested with various parameters with the ultimate goal of providing decision support for cancer riskmanagement.

Reference

Yang C, Yang J, Liu Y and Geng X (2020) Cancer Risk Analysis Based on Improved Probabilistic Neural Network. Front. Comput. Neurosci. 14:58. doi: 10.3389/fncom.2020.00058

Hammo, B. H., Alwidian, J and Obeid, N. (2018). WCBA: weighted classification based on association rules algorithm for breast cancer disease. Appl. Soft Comput.

62, 536–549. doi:

10.1016/j.asoc.2017.11.013

D. A., Decker , Hart, G. R., Roffman, , R., and Deng, J. (2018). A multi- parameterized artificial neural network for lung cancer risk prediction. PLoS ONE 13:e205264. doi: 10.1371/journal.pone.0205264

Iwata, T., Irie, G., Kurashima, T., and Fujimura, K. (2013). Travel route recommendation using geotagged photos. Knowl. Inform. Syst. 37, 37–60. doi: 10.1007/s10115-012-0580-z

Kumar, D Gupta, S., and Sharma, A. (2012). Data mining classification techniques applied for breast cancer diagnosis and prognosis. Indian J. Comput. Sci. Eng. 2,188–195.

Oukid, S., Gagaoua, I. Amrane, M.,, and Ensarl, T. (2018). “Breast cancer classification using machine learning,” in 2018 Electric Electronics, Computer Science, Biomedical Engineerings’ Meeting (EBBT) (Istanbul: IEEE), 1–4. doi:10.1109/EBBT.2018.8391453

Maxim, R andAolin, X.,.(2017). Information-theoretic lower bounds on bayes risk in decentralized estimation. IEEE Trans. Inform. Theory 63, 1580–
1600. doi: 10.1109/TIT.2016.2646342

Grincova, A. Andrejiova, M., (2018).Classification of impact damage on a rubber-textile conveyor belt using nave- bayes methodology. Wear 414–415, 59–
67. doi: 10.1016/j.wear.2018.08.001

Ladhak, F Fakoor, R.,Huber, M. ., Nazi, A (2013). "Using deep learning to enhance cancer diagnosis and classification," in The 30th International Conference on Machine Learning (ICML 2013) (Atlanta, GA), 1–7. Fan,

Che, Y., B., Feng, S., Xie, Y Mao, J., .

(2018). An oil monitoring method of wear evaluation for engine hot tests. Int. J. Adv. Manuf. Technol. 94, 3199–3207.doi: 10.1007/s00170-016-9473-8

Lim, P Kim, H. J., Kim, J. (2018). Towards perfect text classification with wikipedia-based semantic naïve bayes learning.

Mirniaharikandehei, S ,Heidari, M., A. Z., Hollingsworth, A. B. , Khuzani, Danala, G.,., Qiu, Y., et al. (2018). Prediction of breast cancer risk using a machine learning approach embedded with a locality preserving projection algorithm. Phys. Med. Biol. 63:035020. doi: 10.1088/1361-6560/aaa1ca

Handayani, A ,Jamal, A., Ripmiatin, E., Septiandri, A. A., and Effendi, Y. (2018). Dimensionality reduction using pca and k- means clustering for breast cancer

prediction. LontarKomput. 09, 192–201. doi: 10.24843/LKJITI.2018.v09.i03.p08

Neurocomputing 315, 128–134. doi: 10.1016/j.neucom.2018.07.002

Zhang, L and Krause, C. M., (2019).Short-term travel behavior prediction with gps, land use, and point of interest data.Transport.Res. B Methodol. 123, 349–
361. doi: 10.1016/j.trb.2018.06.012

Mayur, S., Jared, W. C., Zaid, R., Thomas, A and ,ARichard. (2019). Sacroiliac joint fusion system for high-grade spondylolisthesis using 'reverse Bohlman technique': a technical report and overview of the literature. World Neurosurg. 124, 331–339. doi: 10.1016/j.wneu.2019.01.041

Lassoued, Y., Russo, G ,Gu, Y, Shorten, R., and Mevissen, M and Monteil, J.,. (2017). "A hidden markov model for route anddestinationprediction,"inIEEE20th

International Conference on Intelligent Transportation Systems (IEEE) (Yokohama), 1–6. doi: 10.1109/ITSC.2017.8317888

Loud, J., and Murphy, J. (2017).Cancer screening and early detection in the 21st century.Semin.Oncol.Nurs. 33, 121–128. doi: 10.1016/j.soncn.2017.02.002

Rajalakshmi, R., and Aravindan, C. (2018). A naïve bayes approach for url classification

with supervised feature selection and rejection framework. Comput.Intell. 34, 363–396.doi:
10.1111/coin. 12158

Bharathi, M., Ezhilarasi, M., Sasikala, S., Senthil, S., and Reddy, M. (2019). Particle swarm optimization based fusion of ultrasound echographic and elastographic texture features for improved breast cancer detection. Australas. Phys. Eng. Sci. Med. 42, 677–688. doi: 10.1007/s13246-019-00765-2