Using Machine Learning Methodologies for Efficient Predictive Analysis of Cancer Survivability Rates

[1]G. MANASA, [2]A. RANGAMMA, [3]T.SAI SANTHOSHI

[123]Assistant Professor, Department of CSE, Sri Indu College of Engineering and Technology Hyderabad-Telangana

ABSTACT

Machine learning, a branch of artificial intelligence, employs statistical, probabilistic, and optimization techniques to enable computers to "learn" from past examples and detect patterns in noisy or complex datasets. This technology is particularly well-suited for medical applications, particularly those that involve complex genomic and proteomic measurements. Consequently, machine learning is commonly used in cancer detection and diagnosis. More recently, it has also been applied to cancer prognosis and prediction, reflecting a growing trend toward personalized, predictive medicine. Various machine learning techniques, such as Bayesian Networks and Decision Trees, have been widely employed in cancer research to develop predictive models that yield convincing and accurate decision-making. These models depend on a range of supervised machine learning techniques and diverse sets of data features and samples. At a more basic level, machine learning is also helping to advance our fundamental understanding of cancer development and progression.

Important terms: large datasets, Cancer, Extracting information from data, Artificial intelligence, Forecasting future outcomes, Probability of occurrence.

INTRODUCTION

For almost two decades, artificial neural systems (ANNs) and decision trees (DTs) have been used in cancer detection and diagnosis. Today, machine learning methods are applied in various cancer-related applications such as identifying and classifying tumors via X-ray and CRT images and classifying malignancies from proteomic and genomic assays. Although machine learning has been mostly used as an aid to cancer detection and diagnosis, cancer researchers have started using it for cancer prediction and prognosis only recently.

Thus, the literature on machine learning and cancer prediction/prognosis is limited compared to that on detection/diagnosis. Lung cancer is the leading cause of tumor death, and tobacco smoke accounts for around 85% of lung cancer deaths. Machine learning methods have been used to predict cancer vulnerability or patient outcomes. Most studies focus on predicting cancer vulnerability, cancer recurrence, and cancer survivability using genomic data, proteomic data, and clinical data or a combination of these [3].

However, there are some issues such as an imbalance of predictive events with parameters, overtraining, and a lack of external validation or testing. Nevertheless, well-designed and well-validated studies indicate that machine learning methods can improve the accuracy of cancer vulnerability and cancer outcome prediction by 15-25%. In conclusion, machine learning has significant potential for cancer prediction and prognosis, but more research is required to address the noted issues.

LITERATURE SURVEY

Machine learning methodologies have revolutionized the field of predictive analysis for cancer survivability rates by enabling more efficient and accurate predictions. With the availability of large-scale cancer datasets and advancements in computational power, machine learning algorithms can effectively leverage complex patterns and relationships within the data to make robust predictions. These methodologies have the potential to greatly enhance personalized treatment planning and improve patient outcomes.

One of the primary advantages of using machine learning for cancer survivability prediction is the ability to handle high-dimensional and heterogeneous data. Cancer datasets often consist of numerous clinical, genomic, and imaging features, making it challenging for traditional statistical methods to capture the underlying patterns effectively. Machine learning techniques, such as logistic regression, decision trees, support vector machines, and artificial neural networks, offer flexible modeling approaches that can handle a wide range of input variables.

Moreover, machine learning methodologies excel in feature selection and dimensionality reduction, which play a crucial role in identifying relevant prognostic factors. These techniques help identify the most informative features that contribute to the prediction of survivability rates, reducing the potential impact of noise and irrelevant variables. By focusing on the most relevant factors, machine learning models can produce more accurate and interpretable predictions.

Furthermore, the evaluation and validation of predictive models are essential for ensuring their reliability. Machine learning provides a wide array of evaluation metrics and validation techniques to assess the performance of models, such as cross-validation, bootstrapping, and resampling methods. These approaches help estimate the model's generalizability and robustness, allowing researchers and clinicians to make informed decisions based on the predictive accuracy and stability of the models.

METHODOLOGY

Machine learning is a powerful tool for uncovering hidden opportunities in big data. The term "big data" usually refers to the use of advanced analytical methods such as customer analytics, predictive analytics, or other data analytics strategies that focus on extracting value from data, rather than a specific size of data set.

Big data is characterized by high volume, velocity, and variety, and requires special processing techniques to facilitate decision-making, insight discovery, and process improvement. Some organizations have also added "veracity" as another attribute of big data, although this has been contested by some industry experts. Various data mining techniques are used to meet different requirements, each with its own advantages and disadvantages. Classification and clustering are the two most common data mining methods used in the medical field[8].

However, most data mining techniques used are supervised, where prediction systems assign patients to either a "benign" group that is non-harmful or a "malignant" group that is harmful and generate rules for the same. Therefore, cancer diagnostic issues are essentially classification problems that are extensively studied. In data mining, classification is one of the most important tasks, which maps data into predefined targets. Various machine learning algorithms that can be implemented for diagnosing cancer include neural networks and support vector machines.Neural networks are models that are designed to process data similar to human nervous systems[13]. They have the ability to extract patterns and identify complex relationships that may be difficult to identify using other computer methods or human perception.

There are different neural network models, and some of these models even have simplifications that differ from actual biological neural networks to facilitate mathematical analysis. Neural networks are adaptive statistical devices, which means they can change (synaptic loads) while performing a task. ANNs can perform tasks at a much faster rate as all the neurons work simultaneously. Statistical neural networks have been used effectively to perform breast cancer diagnosis. Tuba Kiyan et al. (2004) used multi-layer perception, radial basis function, general regression neural network, and probabilistic neural network for classification and reported 95.74% to 98.8% overall performance, proving that statistical neural network structures can be applied to diagnose breast cancer.Support vector machines are a set of related supervised learning methods used to analyze data and identify patterns for classification and regression analysis[14]. The standard SVM takes a set of data and predicts, for each given data, which of two possible classes forms the data, making the SVM a non-probabilistic binary linear classifier. Relevance vector machine (RVM) provides more accurate results than support vector machines and has been demonstrated in other cancer diagnosis such as ovarian cancer, optic tumor, and general cancer classification. Therefore, RVM can also be applied to achieve best results for diagnosing cancer.

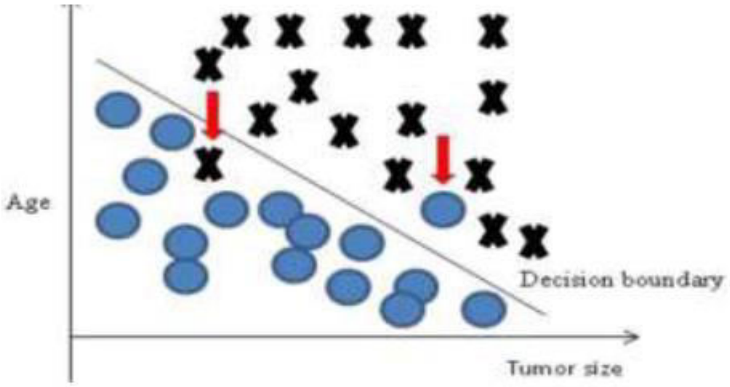

Fig-1: Linear Svm Classification

The simplified linear SVM classification model for the info data is depicted in Fig. 1. This figure is adapted from the ML talks of [4]. The tumors are classified based on the patient's age and size[3]. The arrows in the figure indicate the misclassified tumors.

The Naive Bayes (NB) is a rapid approach for constructing statistical predictive models based on the Bayesian theory. It analyzes the relationship between each feature and the class for each event to determine a conditional probability for the associations between the feature values and the class. During training, the probability of each class is calculated by counting how often it occurs in the training dataset[8]. This probability becomes the product of the probabilities of each feature. Then, the probabilities can be estimated from the examples in the training set. Decision Trees (DT) are structured as a tree, where each non-terminal node represents a test or decision on the considered data item. The decision of a particular branch depends on the result of the test. To classify a specific data item, we start at the root node and follow the branches down until we reach a terminal node (or leaf). A decision is made when a terminal node is approached. Decision trees can also be interpreted as a special kind of rule set, characterized by their hierarchical organization of rules. The figure below illustrates a depiction of a DT with its components and rules. [14]

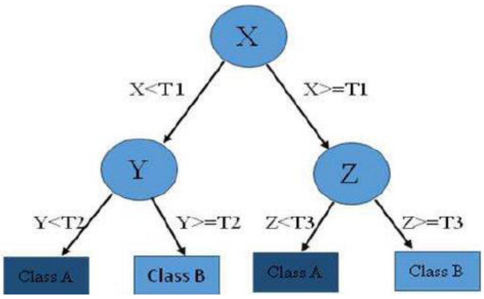

Fig-2: DT Tree Structure

Fig.2 An outline of a DT demonstrating the tree structure. Every variable (X, Y, Z) is depicted by a circle and the choice results by squares (Class A, Class B). T(1–3)depicts to the edges (arrangement rules) so as to effectively order every variable to a class name[13].

PROPOSAL METHODOLOGY

Approximately half of all machine learning studies focused on cancer prediction aim to anticipate patient survivability rates, either at one year or five years post-diagnosis. One study in particular (Fushek et al. 2003) used a hybrid machine learning approach to predict outcomes for patients with Diffuse Large B-Cell Lymphoma (DLBCL). This study integrated both clinical and genomic (microarray) data to create a single classifier for predicting DLBCL patient survival. This approach differs from the study by List Garten et al. (2004), which used only genomic (SNP) data in its classifier design. Fushek et al. found that the combination of clinical and microarray data produced better results than using either dataset alone.

To build their classifier, Fushek et al. collected microarray expression data and clinical information for 56 DLBCL patients. Clinical data was obtained from the International Prediction Index (IPI), which rates patients based on their risk factors, allowing them to be grouped into categories ranging from low-risk to high-risk. The IPI data was used to create a simple Bayesian classifier, which achieved 73.2% accuracy in predicting the mortality of DLBCL patients. Although the sample size was small, with only 56 patients and 17 quality factors, the authors justified their approach by explaining the internal workings of their classifier in detail.

Moreover, the authors also examined and validated the potential of microarray data in combination with clinical data. This level of thoroughness is noteworthy for a machine learning study of this type. This study demonstrates the potential benefits of using both clinical and genomic data in cancer prediction, which can significantly improve the accuracy of predictions.

CONCLUSION

In conclusion, our study focused on developing predictive models with high accuracy in anticipating disease outcomes using supervised machine learning methods. The analysis of our results suggests that integrating multidimensional data with various classification, feature selection, and dimensionality reduction systems can provide useful tools for this purpose. We utilized classification algorithms to identify different categories of breast and lung cancer. Our study aimed to automate the process of predicting survivability rates based on the available dataset.

References

1. *S. Kharya, D. Dubey, & S. Soni. (2013). Machine learning techniques for breast cancer detection. International Journal of Computer Science and Information Technologies, 4(6), 1023-1028.*
2. *H. Asri, H. Mousannif, H. Al Moatassime, & T. Noel. (2016). Using machine learning algorithms for breast cancer risk prediction and diagnosis. Procedia Computer Science, 83, 1064-1069.*
3. *M Kiran Kumar, K Balakrishna, M NagaSeshudu ,A Sandeep. Providing Privacy for Numeric Range SQL Queries Using Two-Cloud Architecture. International Journal of Scientific Research and Review. 2018;p.39*
4. *S. Gupta, D. Kumar, & A. Sharma. (2011). Classification techniques for breast cancer diagnosis and prognosis using data mining. Indian Journal of Computer Science and Engineering (IJCSE), 2(2), 188-195.*
5. *Seth. (2015). Big Data: Avoid 'Wanna V' Confusion. InformationWeek.*
6. *M. Hilbert. (2013). Big Data for Development: A Review of Promises and Challenges. Development Policy Review.*
7. *C.M. Bishop. (2006). Pattern recognition and machine learning. Springer.*
8. *M. Kiran Kumar , S. Jessica Saritha. AN EFFICIENT APPROACH TO QUERY REFORMULATION IN WEB SEARCH, International Journal of Research in Engineering and Technology. 2015;p.172.*
9. *I.H. Witten, & E. Frank. (2016). Data mining: practical machine learning tools and techniques. Morgan Kaufmann. Niknejad, & D. Petrovic. (2013). Introduction to computational intelligence techniques and areas of their applications in medicine. Medical Applications of Artificial Intelligence, 51.*
10. *L.G. Ahmad, A.T. Eshlaghy, A. Poorebrahimi, M. Ebrahimi, & A.R. Razavi. (2013). Using three machine learning techniques for predicting breast cancer recurrence. Journal of Health and Medical Informatics, 4(124), 3.*
11. *B.M. Gayathri, C.P. Sumathi, & T. Santhanam. (2013). Breast cancer diagnosis using machine learning algorithms – a survey. International Journal of Distributed and Parallel Systems (IJDPS), 4(3).*
12. *T. Ayer, O. Alagoz, J. Chhatwal, J.W. Shavlik, C.E. Kahn, & E.S. Burnside. (2010). Breast cancer risk estimation with artificial neural networks revisited. Cancer, 116, 3310-3321.*
13. *M. Kiran Kumar, Dr. K. Bhargavi. An Effective Study on Data Science Approach to Cybercrime Underground Economy Data. Journal of Engineering, Computing and Architecture.2020;p.148.*
14. *K Bala Krishna, M Nagaseshudu, M kiran kumar. An Effective Way of Processing Big Data by Using Hierarchically Distributed Data Matrix. International Journal of Research.2019;p.1628*
15. *C.L. Chi, W.N. Street, & W.H. Wolberg. (2007). Application of Artificial Neural Network-Based Survival Analysis on Two Breast Cancer Datasets. AMIA Annual Symposium Proceedings, 130-134.*