IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Vol. 12, Issue 3, March 2023

DOI: 10.17148/IJARCCE.2023.12335

CLASSIFICATION OF CANCEROUS PROFILE BASED ON TYPE, STAGE AND HETEROGINITY USING MACHINE LEARNING

Dr. Natesh M[1], Manu Prasad[2], Nagendra M R[3], Prajwal[4], Raju M[5]

Associate Professor, Dept. of Computer Science and Engineering, Vidyavardhaka College of Engineering[1]

Dept. of Computer Science and Engineering, Vidyavardhaka College of Engineering[2-5]

Abstract - Cancer is a disease characterized by the abnormal growth and division of cells, which can invade and damage surrounding tissues and organs. It begins when normal cells start to grow and divide uncontrollably due to mutations in genes that control cell growth and division, exposure to certain chemicals or radiation, or a weakened immune system. This survey paper sights to provide an all inclusive analysis ofrecent studies on the classification of cancerous profiles usingmachine learning techniques. The paper focuses on the classification of cancerous profiles based on various parameters, including the type, stage, and heterogeneity of cancer. The paper presents an overview of the different types of cancers and the challenges associated with their diagnosis and treatment. The paper then reviews various machine learning algorithms used in cancer classification and highlights their strengths and limitations.

Keywords - Cancer, Support Vector Machines, Random Forest, k-nearest neighbors (k-NN), LightGBM (LGBM) algorithm, gradient boosted classifier, Convolutional Neural Network (CNN)

I. INTRODUCTION

Cancer is a complex disease that can affect any part of the body. It is caused by the abnormal growth and division of cells, which can invade and damage surrounding tissues and organs. There are many different types of cancer, each with a unique set of qualities, causes, and risk factors.

Cancer begins when normal cells start to grow and divide uncontrollably. This can happen for a variety of reasons, including mutations in genes that control cell growth and division, exposure to certain chemicals or radiation, or a weakened immune system. As the abnormal cells continue to grow and divide, they can form a mass of tissue called a tumor.

Not all tumors are cancerous, however. Tumors that are not cancerous, or benign tumors, do not invade nearby tissues or spread to additional body parts. Malignant tumors, also knownas cancerous tumors, have the ability to invade nearby tissues and organs and spread to other body parts via the lymphatic orblood systems. This process is called metastasis.

There are many different types of cancer, including breast cancer, lung cancer, colon cancer, prostate cancer, and skin cancer. Each type of cancer has its own unique characteristics and can require different treatments.

We discuss different feature selection methods are employedto increase the precision of machine learning models. The survey concludes with a discussion on the future directions of research in cancer classification and the potential impact of machine learning techniques in improving cancer diagnosis and treatment. Overall, this survey provides valuable insights into the current state of research in the classification of cancerous profiles using machine learning techniques.

SVM

Support Vector Machines (SVMs) are a popular machine learning algorithm that can be used for classification and regression analysis. SVMs are particularly useful for classification tasks, where the goal is to predict which class or category a new data point belongs to, based on a set of features or input variables.

In the context of cancer classification, SVMs can be used to analyze data from medical imaging, genetic testing, or other diagnostic tests to predict whether a patient has cancer or not, and what type of cancer they may have. SVMs work by finding the optimal hyper plane that separates the data into different classes, with the maximum margin or distance between each class's closest data points and the hyper plane.

The SVM model is trained using the training set, by feeding it with the input features and corresponding labels (cancer vs. non-cancer) for each data point. The SVM algorithm thenoptimizes the best hyper plane for dividing the two classes, based on the input features.

SVMs have been proven to be efficient in classifying differenttypes of cancer, including breast cancer, lung cancer, and colon cancer. By accurately identifying cancerous tissue or cells, SVMs can help doctors make more informed decisions about treatment options and improve patient outcomes

Random Forest

Random Forest is a popular machine learning algorithm thatcan be used for classification and regression tasks. This particular ensemble learning algorithm combines various decision trees in order to improve the accuracy and stability of the model.

In the context of cancer classification, Random Forest can be used to analyze data from medical imaging, genetic testing, or other diagnostic tests to predict whether a patient has cancer or not, and what type of cancer they may have. Random Forest works by constructing multiple decision trees, where on a subset of each tree's training data of the input features and data points, and the final prediction is made based on the majority vote of all the individual decision trees.

To use Random Forest for cancer classification, researchers typically start by collecting data from patients who have been diagnosed with different types of cancer, as well as fromhealthy individuals who do not have cancer. Then, a training set and a testing set are created from this data.

Random Forest has been demonstrated to be efficient in classifying various cancers, such as breast cancer, lung cancer, and prostate cancer. By accurately identifying cancerous tissue or cells, Random Forest can help doctors make more informed decisions about treatment options and improve patient outcomes

KNN

The k-nearest neighbors (k-NN) algorithm is a machinelearning algorithm used in tasks involving regression and classification. In classification tasks, the k-NN algorithm predicts the class of an input data point by identifying the k closest data points in the training set and classifying the input point based on the most common class of its k nearest neighbors.

In the context of cancer classification, the k-NN algorithm canbe used to predict whether a patient has a cancerous profile or not based on the features of their medical records. For example, a dataset may contain various features such as age, sex, tumor size, and blood markers. By training the k-NN algorithm on this dataset, it can learn to classify new patients based on their similar features. This can be useful for early detection and prevention of cancer, as well as for tailoring treatment plans based on the specific cancerous profile of the patient.

Logistic Regression

Logistic Regression is a widely used statistical method for binary classification, which is a task of predicting the presence or absence of a particular outcome. In the context of cancer classification, logistic regression can be used to predict whether a patient has a cancerous profile or not based on their medical record.

Logistic regression works by modeling the relationshipbetween a set of input features and a binary output variable using a logistic function. The output of the logistic function represents the probability of the positive class (cancerous)given the input features.

To classify cancerous profiles, logistic regression can be trained on a dataset of medical records with known cancerous outcomes. The algorithm learns to assign weights to each feature based on its contribution to the outcome, and then calculates the probability of a patient having a cancerous profile using the logistic function.

DOI: 10.17148/IJARCCE.2023.12335

CNN

The Convolutional Neural Network (CNN) algorithm is a deep learning algorithm that is particularly effective at image recognition tasks. In medical imaging, CNNs can be used to detect and classify cancerous lesions in images such as mammograms, MRIs, and CT scans. CNNs work by processing an image through a series of convolutional layers, which detect patterns and features at increasing levels of abstraction. The output of the convolutional layers is then passed through fully connected layers, which perform the finalclassification task. To classify cancerous profiles using CNNs, medical images of cancerous and non-cancerous tissues are used to train the network.

The CNN learns to distinguish between cancerous and non- cancerous features in the images and can then be used and non-cancerous features in the images and can then be used to classify new images as either cancerous or non-cancerous. This approach has the potential to improve cancer diagnosis accuracy and reduce the need for invasive biopsies, leading to earlier detection and treatment of cancer.

II. BACKGROUND AND RELATED WORK

To address the issue of cancer classification, Wei Luo, LipoWang, and Jingjing Sun proposed a two-step feature selection method using the support vector machine (SVM), modified t- test, and PCA. In gene expression analysis, dimension reduction is accomplished using principal component analysis (PCA). and a modified t-test method is used to choose discriminatory features (genes) that are used to classify cancer. They achieved a classification accuracy rate of 80.4% for cancer profiles. [1]

Dataset pre-processing, clustering using a neural network and classification using a support vector machine (SVM) are performed by Aman Sharma and Rinkle Rani. In an effort to address the classification issue for cancerous profiles, they propose a method, and the results offer a comparison of modelperformance with different sample sizes. The repository for universal genomics of drug sensitivity provided the dataset forthe proposed study (cancer X-gene). With their suggested method, they were able to achieve 85% accuracy and 84% precision for SVM and 83% accuracy and 84.3% precision forNN. [2]

A strategy to create instruments to recognize drug responses in specific patients for precision medicine therapy was put forth by Raihan Rafique, S.M. Riazul Islam, and Julhash U. Kazi. They compared their results with a range of techniques, including SVM, RF, linear regression, KNN, and ANN classifiers. For better outcomes, they combined them withkernel functions and auto-encoders. They discovered that SVM and ANN have least squared error of prediction values of 84% and 83%, respectively (SSE). Consequently, it can be used in drug profiles that predict cancer. [3]

A method to categories cancer types as benign or malignant,as well as whether the disease is curable or not, was proposed by S. Murugan, B. Muthu Kumar, and S. Amudha. They employed Random Forest, Decision Tree, and Linear Regression. The UCI Machine Learning Repository provides the data needed for analysis and breast cancer prediction. Results show that the classification success rate using linear regression is 84.14%, and the prediction success rate using random forest is 88.14%. [4]

Support Vector Machine (SVM) and Random Forest (RF)categorization methods are used in Ashfaq Ahmed, Sultan Aljahdali. To learn, classify, and Data on cancer disease are compared using different kernels and kernel guidelines. The Duke Cancer dataset was taken into account when building the models. An analysis of the outcomes using a confusion matrix (prediction analysis technique). The results with various guidelines are adjusted, and the selection of guidelines for the best classification outcomes is automated. Results can be seen more clearly with SVM than with Random Forest, and SVM produces better accuracy when combined with radial basis function. [5]

The main objective, according to Yaramala Sushma, Vagolu SPrasad Babu, and Vanitha Kakollu, is to offer a method for classifying cancerous profiles using machine learning. Theproposed method addresses the issue of classifying cancerous genomic profiles. The method is based on the idea of using themachine learning algorithms like SVM, kNN, Decision Tree, and Random Forest while implementing validation measures. When the sample size is changed, the result offers a comparison of the model's effectiveness. The performance of the model improves with increasing sample size, demonstrating the model's robustness and adaptability. [6]

The method to identify cancerous patients and classify healthyand cancerous patients was proposed by Poonam Kathale and Snehal Thorat. The Random Forest Classifier technique isused to Categories patients with breast cancer. Performing pre-processing on the input Mammogram image with unwanted elements removed, tumor region segmented using morphological operations, and region highlighted on the original mammogram image. When Mean, GLCM, and Entropy are combined, learning and testing RF classifiers are extremely efficient. [7]

DOI: 10.17148/IJARCCE.2023.12335

Anh Dang suggests analyzing machine learning algorithms by utilizing various models to foretell the likelihood of developing cancer. Cross-validation over the training data set will be used to assess models after training. The models canbe used to test predictions after training. The cross-validation with five folds will be used to test the numerical data set, making the ratio of the train set to the test set 4 to 1. By dividing the original data set into a train set and test set by 80% and 20%, respectively, the image data set will be put to the test. The outcomes of the cross-validation tests for Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Support Vector Machines (SVM), and K-nearest Neighbors (KNN).Decision Tree is over fitting despite having 100% accuracy because the cross-validation scores were only 92–93%. A high cross validation score and excellent accuracy are characteristics of logistic regression. A good model isRandom Forest, which has an accuracy of 96.5% and an average cross-idation score of about 94.5%. Although not as accurate as the other models, Support Vector Machine and K- nearest Neighbors have accuracy levels above 90%.The following chart displays the MobileNetV2 (MB) and Efficient Net (EN)-based image data set's training and testing accuracy. Both models are highly accurate, with a 90% rate. In terms of training accuracy and testing accuracy, MobileNetV2 is marginally superior to EfficientNet. [8]

The clustering-based feature selection (CFS) model, which was proposed by Chandrasegar T. and Sai Brahma Nikhilesh Vutukuri, aims to improve accuracy while minimizing feature dimension. There are three steps involved. Clustering using entropy, BC with all features, and feature entropy partition selection using the best performing classifier. Additionally, we have ML classifiers like Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), and Neural Network (NN) that we can use here (NN). In all the three stages it is observed that Random Forest outperforms all other models. [9]

The effect that various dimensionality reduction techniques have on machine learning models used for cancer prediction was proposed by Md. Faisal Kabir, Tianjie Chen, and Simone A. Ludwig. The dimensionality of the RNA sequencing data was reduced using dimensionality reduction methods like autoencoder, PCA with a kernel, and principal component analysis (PCA). On the original, dimensionally reduced, and cancer-relevant data, two machine learning classifiers—the neural network and the support vector machine—were trained and put to the test machine. The effectiveness of classifiers was evaluated using a variety of metrics, including area under the curve, F-measure, accuracy, precision, recall, and receiver operating characteristic curve. The outcomes demonstrated that dimensionality reduction enhances the classifiers' performance. Additionally, autoencoder outperformed PCA and PCA with a kernel in terms of performance. [10]

III. LITERATURE REVIEW

Table 1: Relevant studies on advantages and disadvantages of Classification of Cancerous profiles

Related Studies	Advantages	Disadvantages	Methodology
[1] “Feature Selection for Cancer Classification Based on Support Vector Machine”	It can be applied on different datasets such as SRBCT dataset,ovarian cancer dataset and leukaemia dataset.	There are changes that can be made in the first step t-test method so thatrequired gene can be eliminated.	PCA, Modified t-test and SVM
[2] “Classification of Cancerous Profiles using Machine Learning”	The results showed a comparisonof the model's performance withdifferent sample sizes. They obtained 85% Accuracy, 84% precision for SVM	They are not able to predict the potential drug targets and drug response.	SVM and Neural Networks
[3] “Machine learning in the prediction of cancer therapy”	They identify drug responses in individual patients for precisionmedicine therapy. They found that SVM and ANN are84% and 83% accuracy and both having least squared error of prediction (SSE).	Utilizes only few parameters and it is limited to drug responsenot about curing the disease.	SVM, RF, Linear Regression, KNN and ANN classifiers.

[4] “Classification and Prediction of Breast Cancer using Linear Regression, Decision Tree and Random Forest”	Using linear regression and randomforest they are able to show classification of cancer (benign or malignant) and whether it is curable or non-curable is also predicted.	Limited to breast cancer and not suitable for othertype of cancer’s.	Linear Regression, Decision Tree and Random Forest.
[5] “Cancer Disease Prediction with Support Vector Machine and RandomForest Classification Techniques”	With the right parameter selectionand the right size training data setswith confusion matrix, their method produces much better results.	With new kernel functions and other classification methods, this research canbe expanded.	Support Vector Machine (SVM) and Random Forest (RF)
[6] “Classification of Cancerous Profiles using Machine Learning Algorithms”	Results showed us the possibility ofusing models based on machine learning for specialists and familiesto accurately and more quickly diagnose cancer	The efficiency of the model is reduced if the training data is very less.	SVM, Decision tree, Random Forest, KNN
[7] “Breast Cancer Detection and Classification”	Uses image processing to determine whether patient has breast cancer or not	Accuracy of randomforestfor breast cancer classification is 84.3% andtakes 9.25 seconds for training and 5.16 seconds for testing.	Median Filtering, Random Forest, Resizing, Mean, GLCM, Entropy
[8] “Cancer Prediction using Machine Learning Algorithms”	Really good accuracy and a high cross validation score are featuresof RF and SVM.	Despite having a decision tree with good accuracy, overfitting is present because the cross- validation scores are onlybetween 84% and 86%. K-nearest Neighbors accuracy is not as good asthe other models.	K-nearest neighbors algorithm (k-NN),Logistic regression, Decision tree classifier, Random Forest, SVM
[9] “Optimized machine learning model using Decision Tree for cancer prediction”	To simplify and improve the performance of ML models of high dimensional data, feature selectionand dimensionality reduction are used.	In clustering based feature selection if number of features is more accuracy is reduced	Decision tree, Random forest, SVM,Neural Network
[10]” A performance analysis of dimensionality reduction algorithms in machine learning models for cancer prediction”	Fewer features mean less complexity, less computation timeand Model accuracy improves dueto less misleading data	We can see that classifiers that used only cancer-relevant features performed significantly worse than those that used all features or reduced features.	Principal component analysis, Autoencoder, Oversampling, SVM, Neural Network

DOI: 10.17148/IJARCCE.2023.12335

IV. CONCLUSION

The common areas of failure that the existing systems face are algorithms that are not able to classify the profile with high accuracy. This may lead to false positives or false negatives, which in this field can mean life or death. The existing systems utilize algorithms such as kNN or Random Forest or decision tree individually for the classification of thedatasets into cancerous profiles, which result in a marginally acceptable accuracy. As for the future work, algorithms such as SVM and Random Forest will be paired and used together on the existing datasets to achieve an accuracy that is much greater when compared to the existing proposals.

REFERENCES

[1] Wei Luo, Lipo Wang , Jingjing Sun, “Feature Selection forCancer Classification Based on Support Vector Machine”, 2019 IEEE

[2] Aman Sharma ,Rinkle Rani, “Classification of Cancerous Profiles using Machine Learning”, Thapar University, Patiala,India - 2020 IEEE

[3] Raihan Rafique , S.M. Riazul Islam and Julhash U. Kazi, “Machine learning in the prediction of cancer therapy”, Lund University, Lund, Sweden - 2022 Elsevier

[4] S. Murugan, B. Muthu Kumar, S. Amudha, “Classification and Prediction of Cancer using Linear Regression, Decision Tree and Random Forest”, Taif, Saudi Arabia - 2017 IEEE

[5] Ashfaq Ahmed K, Sultan Aljahdali, Nisar Hundewale,Ishthaq Ahmed K, “Cancer Disease Prediction with Support Vector Machine and Random Forest Classification Techniques”, Sathyabama University Chennai, India. - 2019 IEEE

[6] Yaramala Sushma, Vagolu S Prasad Babu, Vanitha Kakollu,"Classification of Cancerous Profiles using Machine Learning Algorithms", 2019 IEEE

[7] Poonam Kathale, Snehal Thorat,"Breast Cancer Detection and Classification", 2020 IEEE

[8] Anh Dang,"Cancer Prediction using Machine LearningAlgorithms", Earlham College Richmond, Indiana-2020 IEEE

[9] Chandrasegar T,Sai Brahma Nikhilesh Vutukuri, "Optimized machine learning model using Decision Tree for cancer prediction", SITE, VIT, Vellore, India-2019 IEEE

[10] Md Faisal Kabir, Tianjie Chen, Simone A. Ludwig,"A performance analysis of dimensionality reduction algorithms in machine learning models for cancer prediction", Pennsylvania State University Harrisburg-2022 Elsevier\

[11] Osisanwo F.Y., Akinsola J.E.T., Awodele O., Hinmikaiye J.O., Olakanmi O., Akinjobi J. “Supervised Machine Learning Algorithms: Classification and Comparison".

[12] Y, Mangasarian OL, Wolberg WH. 2000. Breastcancersurvival and chemotherapy: A support vector machine analysis. DIMACS Series in Discrete Mathematics andTheoretical Computer Science.

[13] F.Bray, P. McCarron, and D. M. Parkin, “The changing global patterns of female breast cancer incidence and mortality” Breast Cancer Res.

[14] R.L.Birdwell, D. M. Ikeda, K. D. O'Shaughnessy, and E.

A. Sickles, "Mammographic characteristics of 115 missed cancers later detected with screening mammography and the potential utility of computeraided detection" Radiology.

[15] N.Guler, E. Ubeyli, and I. Guler, "Recurrent neural network employing Lyapunov exponents for EEG signals classifications" Expert systems with Applications.

[16] J.Abonyi and F. Szeifert "Supervised fuzzy clustering for the identification of fuzzy classifiers" Pattern Recognition Letters.

[17] R.Setiono, “Generating concise and accurate classification rules for breast cancer diagnosis” Artificial Intelligence in Medicine.

[18] L.Hadjiiski and B. Sahiner, “Advances in computer-aideddiagnosis for breast cancer. Curr. Opin. Obstet. Gynecol. vol. 18,

[19] J.R.Duda and P. Hart, Pattern Classification and Scene Analysis. John-Wiley, 1973.

[20] D.Specht, “Probabilistic neural networks for classification, mapping or associative memory” in Proc. IEEE Int. Conf. Neural Network.

[21] D.Delen and G. Walker, “Predicting breast cancer survivability: a comparison of three data mining methods” Artificial Intelligence in Medicine.

[22] A.Hong and S. Cho, "Lymphoma cancer classification using genetic programming with SNR features" Lecture Notes on Computer Science.