UDC 575.
https://doi.org/10.2298/GENSR2103297U
Original scientific article

OPTIMISED FEATURE SELECTION FOR EARLY CANCER DETECTION

K.R.UTHAYAN[1*], S.MOHANAVALLI[1], B.NIVETHA[2], S. DHIVYA[1]

[1]Department of Information technology, Sri Sivasubramaniya Nadar College of Engineering, India
[2]Goldman Sachs, Bangalore, India

Uthayan K.R. S.Mohanavalli, B.Nivetha, S. Dhivya (2021). *Optimised feature selection for early cancer detection*. - Genetika, Vol 53, No.3, 1297-1309.

Global Cancer Incidence, Mortality and Prevalence (GLOBOCAN) status report for the year of 2020, suggests the occurrence of 10.0 million cancer deaths and 19.3 million new cancer cases. Clearly, cancer incidence and mortality are rapidly growing worldwide. Also, the leading causes of cancer deaths are found to be lung cancer and breast cancer. Cancer cells are having the probability of spreading to other parts of the body too. Most chronic cancers are not curable, but some can be controlled for a few months or years. Also, there is a possibility of high rate of relapse of the disease. These remissions can be partial or complete. But, if detected early, certain cancers can be treated by surgery, chemotherapy, and radiation therapy. This research work focuses on detecting cancer in its early stage so that right measures can be taken to combat the disease. In this attempt to create a beneficial working model, the combination of Artificial Neural Network (ANN), Convolution Neural Network, Graph based Neural Network with Genetic Algorithm (GA) have proven to be successful. As a proof of concept, we present a combination of feature selection techniques that can effectively reduce the feature set and optimize the classification techniques. The proposed method, when applied on a benchmark dataset, gave a higher accuracy by selecting most relevant 7 features out of 10 with an accuracy of 95.7%. Using Convolution Neural Network, the accuracy improved to 98.3% with optimal hyperparameter tuning.

Key words: Convolution Neural Network, Graph Neural Network, Genetic Algorithm, Logistic Regression, Support Vector Machine

Corresponding author: Uthayan K.R, Department of Information technology, Sri Sivasubramaniya Nadar College of Engineering, India, E-mail: uthayankr@ssn.edu

.

.

INTRODUCTION

Cancer is the deadliest disease a person can ever have. It is caused because of the unregulated cell growth in various parts of the body. The tumor cells can be malignant cancer-causing cells or even benign tumors that don't have any fatal impacts. Thus, addressing this issue at the budding stage can reduce its impact drastically and can save many lives. Recent research on cancer detection is done with the help of X-ray images of the tumor cells (mammography), image processing of the affected areas, infrared thermal imaging of tumor cells, microwave imaging techniques, gene selection and clump features.

Feature selection is the first and most important step in analysis that helps in choosing the relevant features or attributes. It produces a subset feature-set that provides an optimal technique to analyze the dataset accurately. The techniques that can be applied for feature selection are Filter methods, Wrapper methods, Hybrid. The Filter method is independent of the classification algorithm used, while the Wrapper method relies on the performance metrics from the classification algorithm. Hybrid methods include a combination of the former two methods. The steps involved in feature selection are, Subset generation, Subset evaluation, Stopping criteria and Result validation.

The idea of classification is to study the training dataset and learn the pattern. It helps in grouping the data into the target classes based on the training provided. It is a type of supervised learning algorithm, where the features are identified well in advance. There are many classification algorithms, such as Support Vector Machine, Naïve Bayes, Decision Tree, K-nearest neighbor, Fuzzy Clustering, Neural Networks. Among these Artificial Neural Networks (HEPNER *et al*., 1990), Convolution Neural Networks, (LECUN, 2015) and Graph Neural Networks (JIE ZHOU, 2018) can get the insights of the output through the weights imposed on the inputs in different combinations. It has hidden layers to process these inputs and classify the data into appropriate target classes, thus a generalized learning pattern can be observed.

By integrating artificial neural networks with optimal feature selection algorithms, we can provide a solution that could learn and generalize the cancer cell patterns and analyze their behaviors efficiently, based on the training set and derive useful learning patterns, (TAHER *et al.*, 2011).

The research motivation is to combine machine learning and data analytics with a sub field of molecular biology - cancer causing tumor cells - in order to build a detection system that could efficiently classify the malignant cancer cells with minimal features.

MATERIALS AND METHODS

Literature Survey

Many research papers were published, stating the serious impacts of cancer and the approaches in detecting it in early stages. UTHAYAN (2019), proposed an effective diagnosis of cancer by employing microarray gene selection and classification using intelligent dynamic grey wolf optimization technique. A similar approach using an improved grey wolf optimization strategy with enhanced SVM, has been elaborated in the work of WEIYAN *et al*. (2017). MAGLOGIANNIS *et al*. (2009) proposed a solution for detection of breast cancer with the help of clump thickness features through Support Vector Machine based classifiers in comparison with

Bayesian classifiers and Artificial Neural Network for prognosis and diagnosis. SCHÖLKOPF *et al*. (2002) used F-score feature selection technique. ALIČKOVIĆ *et al*. (2017) proposed a similar idea for breast cancer detection using Rotation Forest model with Genetic Algorithm. FATIHAKAY (2009) used a comparative study of four Artificial Neural Networks models, i.e. Back Propagation Algorithm, Learning Vector Quantization, Radial Basis Function Networks and Competitive Learning Network. JANGHEL *et al*. (2010) implemented classification through Memetic Pareto artificial neural network (MPANN) and compared it with Backpropagation algorithm and stated the enhanced performance. RAPPAPORT *et al*. (2005) used the microwave imaging through space-time beamforming to study the early detection of breast cancer. FEAR *et al*. (2002) used the microwave imaging and computes with finite-difference time domain method. TANG (2009) addressed the issue using Computer Aided Diagnosis (CAD). TANG (2009) used mammograms and detection of calcifications, masses, architectural distortion, bilateral asymmetry, image enhancement and image retrieval are performed. ALBA (2007) proposed a comparative study on optimization using Particle Swarm Optimization (PSO) and a Genetic Algorithm (GA) with Support Vector Machine on the selected genes for various types of cancers. WANG *et al*. (2007) used a combination of Fuzzy Neural Network with Support Vector Machine and tried to classify the minimal gene subset efficiently. CHO *et al*. (2002) classified the gene expression data of cancer using ensemble classifier with mutually exclusive features. ARUNA *et al*. (2011) used different classification techniques to study the performance of classifier against the breast cancer dataset. Recently deep learning algorithms have shown remarkable results in detecting tumor tissues from medical images DHIVYA *et al*. (2020).

From the literature, it is evident that the field of medicine requires an innovative approach to predict the fatal disease in a more effective way which will, in turn, simplify the process and yield the result at a faster rate. This would save many lives and spread awareness among people and doctors in an efficient manner. The following are the primary contributions of this paper.

The feature selection techniques that have been used include Genetic Algorithm, Information Gain Ratio, Correlation based, and Gain Ratio based Attribute evaluation and a hybrid of these techniques with Genetic Algorithm, keeping ANN as the base classification technique. Furthermore, the approach has been widened to employ deep learning techniques like Convolution Neural Networks and Graph Neural Network.

The idea of classification is to study the training dataset and learn the pattern. It helps in grouping the data into the target classes based on the training provided. In supervised learning algorithms, features are identified well in advance. There are several classification algorithms such as Support Vector Machine, Naïve Bayes, Decision Tree, K-Nearest Neighbor, Fuzzy Clustering, Neural Networks, NAROUI (2020). Among these, Artificial Neural Networks can get the insights of the output through the weights imposed on the inputs in different combinations. It has hidden layers to process these inputs and classify the data into appropriate target classes, thus a generalized learning pattern can be observed.

By integrating neural networks with optimal feature selection algorithms, we can provide a solution that can learn and generalize the cancer cell patterns, analyze their behaviors efficiently based on the training set and derive useful learning patterns. Using the performance metrics, these results are compared and evaluated. This complete process is explained in the form of flow diagram using Figure 1.

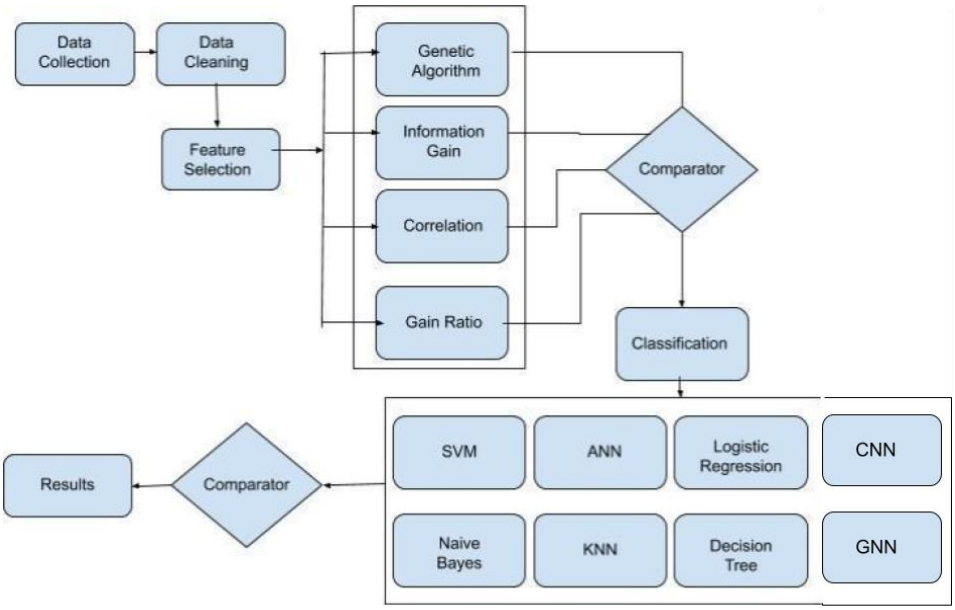

Figure1. Flow diagram

Data Collection

Data collection is the process of collecting relevant data and identifying features that would be suitable to derive the solution. While performing data collection, we must understand the various root causes and ensure the relevance of the data to the stated problem.

Data Cleaning

The act of finding and correcting (or removing) incorrect or inaccurate data from a group of records, as well as changing, replacing, or deleting the unwanted data, is known as data cleaning. Data cleansing can be done in real time using data wrangling tools or in batches using scripting. It involves different methods of removing the missing values, pruning the unclear data, removing the outliers, noise etc.

Feature Selection

Feature selection is the primary and most important step in analysis that helps in choosing the relevant features or attributes. It produces a subset feature-set that provides an optimal technique to analyze the dataset accurately. The techniques that can be applied for feature selection are Filter methods, Wrapper methods, Hybrid. Filter method is independent of the classification algorithm used, while the Wrapper method relies on the performance metrics from the classification algorithm. Hybrid methods include a combination of the former two methods.

Subset generation, Subset evaluation, stopping criteria and result validation are the phases involved in feature selection.

Genetic Algorithm

The genetic algorithm (GA) is a generic adaptive optimization search approach based on Darwin's idea of natural selection - "Survival of the Fittest" and genetics in biological systems. It's a promising replacement for traditional heuristic approaches. GA employs a population, also known as the set of chromosomes, of possible solutions. It iterates through a set of calculations. GA iterates across populations of alternative solutions represented by a chromosome, i.e., a solution to the problem, until acceptable results are obtained. After the evaluation process, a fitness function determines the quality of a solution. The crossover and mutation functions are the two most main operators that influence the fitness value.

Information Gain

Information Gain Attribute evaluation is a feature selection technique that evaluates the importance of an attribute using its entropy value. Information Gain is also known as mutual information.

InfoGain (Class, Attribute) =H (Class) - H (Class | Attribute)→Eq (1) where H is the Information entropy.

Correlation Feature Selection

The Correlation Feature Selection (CFS) analyses the subsets of features based on the hypothesis that the best features are the ones which are highly correlated with the dependent variable and uncorrelated/ collinear with the other independent variables.

Gain Ratio

Information gain ratio is a ratio of calculated information gain to the intrinsic information. It is used to reduce bias towards multi-valued attributes by taking the number and size of branches into account while choosing an attribute.

Classification

The idea of classification is to study the training dataset and learn the pattern. It helps in grouping the data into the target classes based on the training provided. It is a type of supervised learning algorithm, where the features are identified well in advance. There are many classification algorithms, such as Support Vector Machine, Naïve Bayes, Decision Tree, K-Nearest Neighbor, Fuzzy Clustering, Neural Networks [1, 2, 5]. Among these, Artificial Neural Networks can get the insights of the output through the weights imposed on the inputs in different combinations. It has hidden layers to process these inputs and classify the data into appropriate target classes, thus a generalized learning pattern can be observed.

Artificial neural networks

Among the options available for classification, in medicine, the most widely used models are logistic regression (LR) and artificial neural networks (ANN). Artificial neural networks are

biologically inspired networks—inspired by the human brain and its neuronal arrangement. The decision-making process of ANN is more holistic, based on the sum of all input patterns.

Support vector machines

Support vector machines solve a constrained quadratic optimization problem by dividing data sets with a hyperplane to create optimal separating borders between them. The degree of nonlinearity can be changed by employing different kernel functions, making the model more flexible. The disadvantage of a support vector machine is that the classification result is purely dichotomous (in which there are only two possible outcomes) and there is no likelihood of class membership.

K-Nearest Neighbor

The K-Nearest Neighbor approach is different from the other methods considered here in that it uses the data directly for classification rather than first building a model. The model's single adjustable parameter is K, which is the number of nearest neighbors to consider for estimating class membership: P(y|x) is simply the ratio of members of class y among the K-nearest neighbors of x. The model can be made flexible by changing K.

Decision Tree

Using this technique, the data is repeatedly split into sets according to a condition which maximizes the separation of the data, resulting in a tree-like structure. The most common condition used for splitting is information gain.

Logistic regression

Logistic regression is one of the statistical methods to analyze a dataset which contains one or more independent variables to determine an out- come. This outcome is measured with a distinguishing parameter. Here the dependent variable is binary i.e., it only contains data encoded as 1 (TRUE) or 0 (FALSE), DREISEITL *et al*. (2002).

Naïve Bayes

Naïve Bayes classifier is a powerful algorithm for the classification. It can handle any size of dataset ranging from small to large. It is supported by the concepts of prior and conditional probability.

Graph Neural networks

Graph Neural networks (GCNs) have shown a remarkable evolution in the generalization of neural networks on graph structured data. The GCN are typically a multi-layerperceptron (MLP) with an adjacent matrix. The GCN are two types of spectral based models and spatial based models. The representation of a given data into a graph provides more insights about the structural content and their relations. These graph representations are done in a low-dimensional Euclidean space using many techniques based on KNN graphs and other embedding methods. In this work, the impact of GCN is analyzed on regular structure data, WDBC data collected from UCI repository.

Convolutional neural networks

The convolutional neural networks (CNN) have remarkably shown improvisation in tasks such as object detection, segmentation, and classification. The CNN contains two phases: the feature extraction phase and the classification phase. Under the feature extraction phase, it includes convolutional and pooling layers. In the convolutional layer, several filters or kernels are convolved with the input matrix to obtain the feature maps. The filter or kernel is used to infer the underlying pattern in the given data. The pooling layer is used for down sampling the obtained feature maps by reducing the size of spatial representation. As a result of these two layers, we obtain the global features. To introduce non-linearity, the activation function used in the convolutional layer is ReLU(Rectified linear units) in order to assign negative values as 0, thereby preventing many nodes from being a part of the training phase. The second phase involves the classification process through dense layers, which acts as a traditional neural network. The feature maps that are obtained are flattened and are fed to a dense layer which is composed of nodes to process the inputs and their weights. The activation function used in the last layers is the softmax which converts the given input vectors into class probabilities.

RESULTS AND DISCUSSION

To detect cancer cells, we have experimented with two datasets i.e., Wisconsin-Breast Cancer datasets - Original and Diagnostics. Based on cell features the original database Wisconsin breast cancer diagnosis (WBCD) was used to classify a tumor as benign or malignant. Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Clump Thickness, Single Epithelial, Cell Size Bare Nuclei, Bland Chromatin, Normal Nucleoli, Marginal Adhesion, and Mitoses are among the 10 attributes in the dataset, which also includes the class distribution Benign and Malignant. There are 699 cases in the database, 458 of which are benign and 241 of which are malignant. There are approximately 16 attribute values that are missing. As a result, we haven't taken them into account. In the Diagnostic dataset, there are 569 instances with 32 attributes: ID number, Diagnosis (M = malignant, B = benign), and so forth. For each cell nucleus, ten real-valued features have been computed: a) the radius (mean of distances from center to points on the perimeter) b) texture (standard deviation of gray-scale values) c) area d) perimeter e) smoothness (local variation in radius lengths) f) compactness (perimeter^2 / area - 1.0) g) concavity (severity of concave portions of the contour) h) concave points (number of concave portions of the contour) i) symmetry j) fractal dimension ("coastline approximation" - 1) along with their mean, standard deviation and worst values.

The process of data cleaning involves the removal of missing data as there were very few records with missing values and their removal did not affect performance of prediction.

The cleaned dataset has been passed over feature selection techniques like Genetic Algorithm, Information Gain, Correlation based attribute selection and Gain Ratio based attribute selection individually and then classification is done over the selected features as shown in Figure 1.

The feature subsets that have been selected from Genetic Algorithm includes size uniformity, shape uniformity, marginal adhesion, epithelial size, bare nucleoli, normal nucleoli, and bland chromatin. Other techniques such as Information Gain, Correlation based attribute selection and Gain Ratio based attribute selection produced the following results which are

mentioned in Table 1.

Table1.Feature Subsets from different Feature Selection Techniques

Feature Selection Technique	No. of Features	Feature Subset	Percentage of correctly classified instances
InfoGain	7	{1,2,3,5,6,7,8}	95.7143
Correlation	7	{1,2,3,4,6,7,8}	96.1905
Gain Ratio	7	{2,3,5,6,7,8,9}	95.2381

Furthermore, a combination of these feature selection techniques has been applied and the following subsets have been generated which are mentioned in Table 2.

Table 2. Feature Subsets from combination of Feature Selection Techniques

Feature Selection Technique	No. of Features	Feature Subset	Percentage of correctly classified instances
GAwithInfoGain	5	{1,2,5,6,7}	95.7143
GAwithCorrelation	5	{1,2,5,6,7}	95.7143
GAwithGainRatio	5	{2,5,6,7,8}	96.67

These feature subsets have been worked upon by different classification techniques like SVM, Logistic Regression, ANN, CNN, GNN, Naïve Bayes, K-Nearest Neighbor and Decision Tree using Python and its results have been mentioned in the below Figure 2.

These graphs show a cumulative performance comparison of the feature subsets from the combination of feature selection techniques (Genetic Algorithm and Gain Ratio) against the classifiers like Artificial Neural Networks, Naïve Bayes, Support Vector Machine, Decision Tree, K -Nearest Neighbor and Logistic Regression. It has been observed that ANN provides better results with 96.67% against the given feature subsets, which is the best in comparison to other classifiers. In the case of the original dataset, higher accuracy has been obtained by selecting 7 features out of 10 with an accuracy of 95.7%. But in the diagnostic dataset, the features have been reduced by almost half, to 14 from 31 without any compensation in the results and the efficiency has been improved to 96.67% using the combination of Genetic Algorithm.

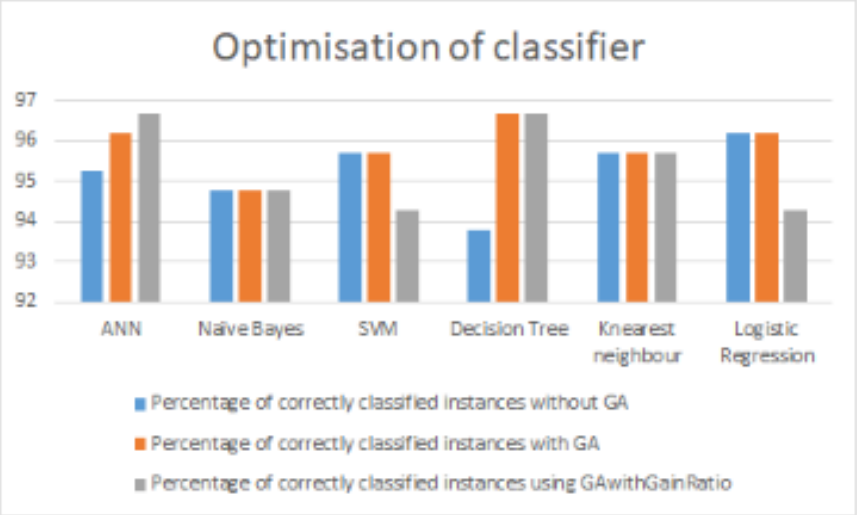

Figure 2. Comparison of Percentage Accuracy from different classification techniques over the feature subsets

Recent research works shows that Deep Learning techniques contribute to improved tumor detection. Classification of WDBC data set was experimented with Graph based Neural Networks.To work with the graph network, the data is preprocessed and transformed to low-dimensional representation space. The nearest neighbor of the data points is found using the Euclidean distance. Then the data points are aggregated for train, validation, and test. The graph representation is fed to a graph convolution network which is composed of two layers with activation function as Rectified linear Unit (ReLU). Adam optimizer is employed with a learning rate of 3e-4 and the weights are adjusted after epochs. The model achieved an accuracy of 94.3% and precision of 92.1%. The results are illustrated in Table3.

Table 3. Performance for GCN

Score Type	Average Score
Accuracy	94.3%
Sensitivity	92.1%
Recall	77.3%
Specificity	93.1%

The datasets WDBC and WBCD involve the cytological features that were obtained as a result of fine needle aspiration from breast masses. Nonetheless, both the datasets are structured data. There are nearly 39 features, and are represented as a vector, hence the traditional CNN that comprises 2D data for the image could not be employed for this data. Hence, the convolutional and pooling layers are employed using 1D space. A CNN should typically include at least one

convolutional layer. The convolutional layer employed is 1D, where the inputs are convolved with the filters and the obtained feature maps are reduced using maxpooling function in the pooling layers. Three convolutional layers followed by maxpooling layer is used to build the feature extraction phase. To avoid overfitting, dropout is employed, where some nodes are prevented to take part in the learning process. The obtained feature maps are flattened and are fed to the first dense layer. In the second dense layer, the softmax function is used as the activation function. The loss is calculated using the binary cross entropy which is further minimized using Adam optimizer with a learning rate of 0.001. The weights are adjusted after each epoch and for 50 epochs the model yielded 97.9% of accuracy along with a precision of 97.1% and specificity of 98.3%. The results are tabulated in Table 4. The overall comparison is illustrated in Fig 3.

Table 4. Performance for CNN

Score Type	Average Score
Accuracy	98.3%
Sensitivity	97.1%
Recall	89.0%
Specificity	97.9%

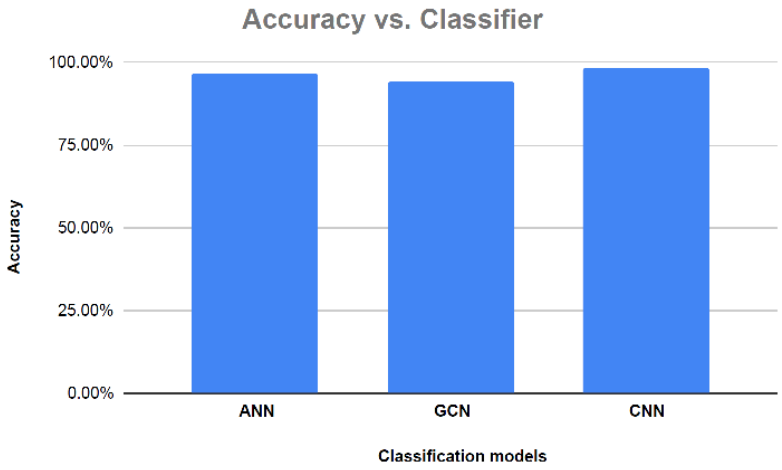

Figure 3. Comparison of best performing classification models

CONCLUSION

Thus, an early tumor diagnosis method to classify the breast cancer dataset by employing ANN has been developed and its performance has been further improved by using a suitable feature selection technique (i e. Genetic Algorithm (GA) or a combination of GA with other filters). Also, it was observed that Genetic Algorithm yields better results irrespective of the size of the dataset, which leads to an inference that with more the generations and epochs, higher classification performance could be attained. On comparing the results of ANN with other classifiers like Naïve Bayes, SVM, Decision Tree, K- Nearest Neighbor and Logistic Regression, it could be seen that higher improved performance has been obtained, based on various metrics. Also in this work, we have employed graph techniques to find graph structure of the given dataset as they do not possess graph representations explicitly. The results of the GCN can be improved by optimizing the graph representation using better techniques as the GCN completely relies on these graph representations. Hence, learning the graphs and training them simultaneously may improve the quality of the graph, thereby increasing the GCNs performance. However, CNNs can extract the underlying patterns and classify the given mass lesions with utmost accuracy. As future works, the GCN and CNN models can be used to validate image datasets.

Received, January 28th, 2021
Accepted September 28th, 2021

REFERENCES

AKAY, M.F. (2009): Support vector machines combined with feature se- lection for breast cancer diagnosis. Expert systems with applications, *36*(2): 3240 – 3247.

ALBA, E. (2007): Gene selection in cancer classification using PSO/SVM and GA/SVM hybrid algorithms, Proc. in IEEE *Congress on*. Evolutionary Computation.

ALIČKOVIĆ, E. and A., SUBASI (2017): Breast cancer diagnosis using GA feature selection and Rotation Forest. Neural Computing and Applications, *28*(4): 753-763.

ARUNA, S., S.P., RAJAGOPALAN, L.V., NANDAKISHORE (2011): Knowledge based analysis of various statistical tools in detecting breast cancer, Computer Science & Information Technology, *2*: 37–45.

CHO, S.B. and J., RYU (2002): Classifying gene expression data of cancer using classifier ensemble with mutually exclusive features. Proc. in IEEE, *90*(11): 1744–1753.

DHIVYA, S., R.J., ANJALI, S., MOHANAVALLI, N., SRIPRIYA, K., SRINIVASAN (2020): Investigations of Shallow and Deep Learning Algorithms for Tumor Detection, Roc. In *2020* IEEE-HYDCO, IEEE: 1-5.

DREISEITL, S., L., OHNO-MACHADO (2002): Logistic regression and artificial neural network classification models: a methodology review. J. Biomed. Informatics, *35*(5-6): 352–359.

FEAR, E.C. (2002): Confocal microwave imaging for breast cancer detection: Localization of tumors in three dimensions. IEEE Transactions on Biomedical Engineering, *49*(8): 812–822.

HEPNER, G. (1990): Artificial neural network classification using a minimal training set- Comparison to conventional supervised classification. Photogrammetric Engineering and Remote Sensing, *56*(4): 469–473.

JANGHEL, R.R, A., SHUKLA, R., TIWARI, R., KALA (2010): Breast cancer diagnosis using artificial neural network models, Proc. in Information Sciences and Interaction Sciences, China.

KOSMAS, P., C.M., RAPPAPORT (2005): Time reversal with the FDTD method for microwave breast cancer detection, IEEE Transactions on Microwave Theory and Techniques, *53*(7): 2317 – 2323.

LECUN, Y., Y., BENGIO, G., HINTON (2015): Deep learning. Nature, *521*(7553): 436-444.

MAGLOGIANNIS, I., E., ZAFIROPOULOS, I., ANAGNOSTOPOULOS (2009): An intelligent system for automated breast cancer diagnosis and prognosis using SVM based classifiers, Applied intelligence, *30*(1): 24-36.

NAROUI, R. M. R., G., KEYKHA, J., ABBASKOOHPAYEGANI, R., RAFEZI (2020): Machine learning approaches to classify melon landraces based on phenotypic traits. Genetika, *52*(3): 1021-1029.

SCHÖLKOPF, B. and A.J., SMOLA (2002): Learning with kernels: support vector machines, regularization, optimization, and beyond, MIT press.

TAHER, F. and R., SAMMOUDA (2011): Lung cancer detection by using artificial neural network and fuzzy clustering methods, Proc. in IEEE GCC.

TANG, J. (2009): Computer-aided detection and diagnosis of breast cancer with mammography: recent advances", IEEE Transactions on Information Technology in Biomedicine, *13*(2): 236–251.

UTHAYAN, K.R. (2019): A novel microarray gene selection and classification using intelligent dynamic grey wolf optimization. Genetika-Belgrade, *51*(3): 805-828.

WANG, L., F., CHU, W., XIE (2007): Accurate cancer classification using expressions of very few genes, IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), *4*(1): 40–53.

YAN, W., N., NI, D., LIU, H., CHEN, M., WANG, Q., LI, X., CUI (2017): An improved grey wolf optimization strategy enhanced SVM and its application in predicting the second major", Mathematical Problems in Engineering. Hindawi, pp. 1-12.

ZHOU, J., G., CUI, S., HU, Z., ZHANG, C., YANG, Z., LIU, *et al*. (2020),"Graph neural networks: A review of methods and applications", AI Open, *1*: 57-81.

OPTIMIZOVANI IZBOR KARAKTERISTIKA ZA RANO OTKRIVANJE RAKA

K.R.UTHAYAN[1*], S.MOHANAVALLI[1], B.NIVETHA[2]and S. DHIVYA[1]

[1]Departman za informacione tehnologije, Sri Sivasubramaniya Nadar Koledž za inžinjerstvo, Indija
[2]Goldman Sachs, Bangalore, Indija

Izvod

Globalni izveštaj o pojavi, mortalitetu i rasprostranjenosti raka (GLOBOCAN) za 2020. godinu ukazuje na pojavu 10,0 miliona smrtnih slučajeva od raka i 19,3 miliona novih slučajeva raka. Jasno je da učestalost raka i smrtnost brzo rastu širom sveta. Takođe, utvrđeno je da su vodeći uzroci smrti od raka rak pluća i rak dojke. Ćelije raka imaju verovatnoću da se šire i na druge delove tela. Većina hroničnih karcinoma nije izlečiva, ali neki se mogu kontrolisati nekoliko meseci ili godina. Takođe, postoji mogućnost visoke stope relapsa bolesti. Ove remisije mogu biti delimične ili potpune. Ali, ako se rano otkriju, određeni karcinomi se mogu lečiti operacijom, hemoterapijom i terapijom zračenjem. Ovaj istraživački rad se fokusira na otkrivanje raka u ranoj fazi kako bi se mogle preduzeti prave mere za borbu protiv bolesti. U ovom pokušaju stvaranja korisnog radnog modela, kombinacija veštačke neuronske mreže (ANN), konvolucione neuronske mreže, neuronske mreže zasnovane na grafu sa genetskim algoritmom (GA) pokazala se uspešnom. Kao dokaz koncepta, predstavljamo kombinaciju tehnika selekcije karakteristika koje mogu efikasno smanjiti set karakteristika i optimizovati tehnike klasifikacije. Predložena metoda, kada je primenjena na referentni skup podataka, dala je veću tačnost odabirom najrelevantnijih karakteristika od 10 sa tačnošću od 95,7%. Korišćenjem konvolucione neuronske mreže, preciznost je poboljšana na 98,3% uz optimalno podešavanje hiperparametara.

Primljeno 28.I.2021.
Odobreno 28. IX. 2021.