CLASSIFICATION OF CERVICAL CANCER IN WOMEN USING CONVOLUTIONAL NEURAL NETWORK

M. Ramkumar[1], R. Manikandan[2], M. Punithavalli[3], V.S. Akshaya[4] and Shanmugaraj Madasamy[5]

[1]*Department of Computer Science and Engineering, Gnanamani College of Technology, India*
[2]*Department of Computer Science, The Quaide Milleth College for Men, India*
[3]*Department of Mechanical Engineering, Rathinam Technical Campus, India*
[4]*Department of Computer Science and Engineering, Sri Eshwar College of Engineering, India*
[5]*Department of Computer Science, Cork Institute of Technology, Ireland*

Abstract

Cervical cancer is regarded as a serious threats to humanity, globally and this is a vital disease with huge spreading of virus that affects the health of humans. The virus is spreading at a rapid rate through mosquitoes that even may kill the one who is affected with cervical cancer. In this paper, we develop a quick response system that certainly finds the disease through a faster validation process. The study uses Convolutional Neural Network (CNN) as a deep learning model that classifies and predicts the condition or the infection status of a patient. The study uses a pre-processing model and a feature extraction model to prepare the image datasets for classification. The simulation is conducted to validate the effectiveness of the model over cervical cancer image datasets i.e. the blood samples of humans. The validation shows that the proposed method effectively classifies the patients in a faster manner than the other deep learning models.

Keywords:

Machine Learning, Cervical cancer, Classification, Diagnosis

1. INTRODUCTION

The human cervix is coated by a thin cell layer of tissue. When a cell becomes a malignant cell which can grow quickly and create a tumour, we term cervical cancer. This type of cell changes. This cancer is treated in its early stages. It may be treated. In many circumstances, a screening method and a biopsy are used to diagnose. These approaches of image processing help to detect the extent of cancer growth. Cervical cancer is the fourth most frequent cause of cancer mortality in women [1]. Medical image processing and smart systems are vital for malignant cell analysis. When new approaches are discovered, they can be obtained at a reasonable cost and in less time-consuming methods.

Real time, precision and human-free interaction: We have developed a complete medical diagnosis system that can be implemented in real time, accurately and without human touch using state of the art machine learning technology and wireless communication technologies.

Predictive modelling of classification is the act of taking a newly observed sample and allocating it, by means of a trained model, to one of the specified categories, also known as labels. This categorises data in order to create the foundation for future applications of computer vision, such as detection, location and segmentation [1]. Although this activity is commonly considered to be simple for humans, the task for a computer based system is significantly more complicated, both as a result of problems resulting from object variability and because of very high class variability in the numerous types of objects [2].

In the realm of modern technology, researchers and engineers around the world use CNNs to address various computer vision classification challenges. Researchers generated incredibly low classification-related error rates in a variety of photos using neural networks [3] [4]. CNN still has certain limitations and weaknesses in spite of its excellent results. CNNs acquire the characteristics of every layer; this starts with the discovery of borders, then geometric shapes and finally real things. In particular, this applies to information concerning the spatial links between the characteristics, which is hard to access [23]-[25] (perspective, size and orientation). Images of properties or samples from the same photos in a different direction seem to easily deceive them. Many approaches to resolving this problem are possible, but this takes outstanding training in all possible approaches. Furthermore, it takes time and computer resources to resolve this problem [5]-[7].

As a result, population screening in underdeveloped nations is still not widely available. Parts of the population-wide screening in underdeveloped countries are not easily accessible because of the long and complex nature of the manual detection of aberrant cells from cervical cytology samples [4]. Not only can the actual diagnosis be mistaken, but even well-trained doctors sometimes miss key findings [4]. In the last few decades, automation-aided reading systems have been developed to improve efficiency and enhance general availability to address these challenges. The aforesaid automated image analysis approaches are based on automated image analytics [4] – [6], that are automatically used for further evaluation by the cytoscreener or cytopathologist to identify probable issue zones from a given cervical cytology slide. Under the World Health Organisation's classification system, patients with premalignant cell dysplastic changes have four different phases, with the first of mild, moderate, severe and in situ (CIS) carcinoma [7]. The most frequent skin lesions are categorised into two categories: those without a pronouncement other than lesions, which are indicators of mild stage HPV infection and those with a greater risk of developing cancer if not recognised and treated. In order to avoid pre-cancer cell formation, dysplastic changes in the early stage are crucial. Fine grain classification is crucial when dealing earlier in the course of the illness, although previous research analysing cervical cells based on automated image analysis focused instead on classification of normal and malnormal cervical cells [4], [6], [9] – [14].

In this report, we design a quick response system, with a speedier validation process that certainly finds the sickness. The study uses CNN as a learning machine to classify and forecast the patient's state or infection status. The study employs an image dataset for classification with a pre-processing model and a feature extraction technique. The image datasets, i.e. blood

samples of people, are performed to evaluate the efficacy of the model against cervical cancer. The validation reveals that the strategy suggested classifies patients successfully faster than alternative models of deep learning.

2. RELATED WORKS

The human cervix is coated by a thin cell layer of tissue. When a cell becomes a malignant cell which can grow quickly and create a tumour, we term cervical cancer. This type of cell changes. This cancer is treated in its early stages. It may be treated. In many circumstances, a screening method and a biopsy are used to diagnose. These approaches of image processing help to detect the extent of cancer growth. Cervical cancer is the fourth most frequent cause of cancer mortality in women [1]. Medical image processing and smart systems are vital for malignant cell analysis. When new approaches are discovered, they can be obtained at a reasonable cost and in less time-consuming methods.

Real time, precision and human-free interaction: We have developed a complete medical diagnosis system that can be implemented in real time, accurately and without human touch using state of the art machine learning technology and wireless communication technologies.

Predictive modelling of classification is the act of taking a newly observed sample and allocating it, by means of a trained model, to one of the specified categories, also known as labels. This categorises data in order to create the foundation for future applications of computer vision, such as detection, location and segmentation [1]. Although this activity is commonly considered to be simple for humans, the task for a computer based system is significantly more complicated, both as a result of problems resulting from object variability and because of very high class variability in the numerous types of objects [2].

In the realm of modern technology, researchers and engineers around the world use CNNs to address various computer vision classification challenges. Researchers generated incredibly low classification-related error rates in a variety of photos using neural networks [3] [4]. CNN still has certain limitations and weaknesses in spite of its excellent results. CNNs acquire the characteristics of every layer; this starts with the discovery of borders, then geometric shapes and finally real things. In particular, this applies to information concerning the spatial links between the characteristics, which is hard to access [23]-[25] (perspective, size and orientation). Images of properties or samples from the same photos in a different direction seem to easily deceive them. Many approaches to resolving this problem are possible, but this takes outstanding training in all possible approaches. Furthermore, it takes time and computer resources to resolve this problem [5]-[7].

As a result, population screening in underdeveloped nations is still not widely available. Parts of the population-wide screening in underdeveloped countries are not easily accessible because of the long and complex nature of the manual detection of aberrant cells from cervical cytology samples [4]. Not only can the actual diagnosis be mistaken, but even well-trained doctors sometimes miss key findings [4]. In the last few decades, automation-aided reading systems have been developed to improve efficiency and enhance general availability to address these challenges. The aforesaid automated image analysis approaches are based on automated image analytics [4] – [6], that are automatically used

for further evaluation by the cytoscreener or cytopathologist to identify probable issue zones from a given cervical cytology slide. Under the World Health Organisation's classification system, patients with premalignant cell dysplastic changes have four different phases, with the first of mild, moderate, severe and in situ (CIS) carcinoma [7]. The most frequent skin lesions are categorised into two categories: those without a pronouncement other than lesions, which are indicators of mild stage HPV infection and those with a greater risk of developing cancer if not recognised and treated. In order to avoid pre-cancer cell formation, dysplastic changes in the early stage are crucial. Fine grain classification is crucial when dealing earlier in the course of the illness, although previous research analysing cervical cells based on automated image analysis focused instead on classification of normal and malnormal cervical cells [4], [6], [9] – [14].

In this report, we design a quick response system, with a speedier validation process that certainly finds the sickness. The study uses CNN as a learning machine to classify and forecast the patient's state or infection status. The study employs an image dataset for classification with a pre-processing model and a feature extraction technique. The image datasets, i.e. blood samples of people, are performed to evaluate the efficacy of the model against cervical cancer. The validation reveals that the strategy suggested classifies patients successfully faster than alternative models of profound learning.

3. PROPOSED METHOD

The prediction model for cervical cancer occurrences operates in four steps. The four-phase workflow is seen in the Fig.1.
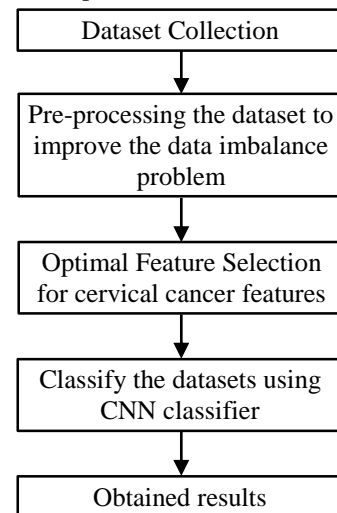
```
┌─────────────────────────────┐
│     Dataset Collection      │
└─────────────────────────────┘
              │
              ▼
┌─────────────────────────────┐
│ Pre-processing the dataset to│
│  improve the data imbalance  │
│          problem             │
└─────────────────────────────┘
              │
              ▼
┌─────────────────────────────┐
│  Optimal Feature Selection   │
│  for cervical cancer features│
└─────────────────────────────┘
              │
              ▼
┌─────────────────────────────┐
│  Classify the datasets using │
│        CNN classifier        │
└─────────────────────────────┘
              │
              ▼
┌─────────────────────────────┐
│       Obtained results       │
└─────────────────────────────┘
```

Fig.1. Proposed prediction model

- **Phase 1**: Real time data collection for cervical cancer disease
- **Phase 2**: Data Pre-processing.
- **Phase 3**: Feature selection by weighted score
- **Phase 4**: Classification using ANN

Medical (clinical) data refers to the personal information that is involved or part of a clinical trial scheme with routine patient treatment. This covers digital copies of personal care data [19],

electronic health reports. The collection, storage, processing, analysis or dissemination of medical data is a broad volume [20]-[22]. Data samples from Cervical cancer Disease are obtained from different health centers across India as part of the real-time data collection process.

3.1 PREPROCESSING

Due to the enormous volume of diagnostic information from the different sources, it is possible to cope with certain problems, including lost values, noise, mislabeling, high dimensionality and imbalances. This proposed research applies the preprocessing task to manage imbalanced data, which is the Symmetric Minority Over Sampling Technique (SMOTE). This improves the accuracy of the classification of different traditional classifiers. These findings give improved results over the balanced dataset. The results are better. For the balanced dataset, the following formula is used:

$$\left(A'_{new_1}, A'_{new_2}\right) = \left(A_{11}, A_{12}\right) + rand\left(0-1\right)*\left(A_{new_1}, A_{new_2}\right) \quad (1)$$

where

$A_{new_1} = A_{21}\text{-}A_{11}$ and $A_{new_2} = A_{22}\text{-}A_{12}$

A_{11} - first attribute of first instance,

A_{21} - first attribute of second instance,

A_{12} - second attribute of first instance and

A_{22} - second attribute of second instance.

rand(0-1) - random number.

Following the pre- mission, conventional feature selection methods like Genetic Algorithm (GA), Correlation based Feature Selection (CFS) and Particle Swarm Optimization (PSO) are evaluated by using the SMOTE over imbalanced data sets for the pre-processor. Afterwards, the findings are analyzed using the balanced dataset.

3.2 FEATURE EXTRACTION

In order to calculate the degree of complexity of the knowledge quality, 'Entropy' proposed by Shannon is used. The uses entropy-weighted methods in our proposed approach to derive objective weights of associated attributes. The proposed solution is given in following steps:

Step 1: Create decision matrix with the weighted score.

Step 2: Normalize the decision matrix to pre-process the data.

Step 3: Generate the weighted score matrix.

Step 4: Optimize weights using entropy weighted method.

Step 5: Compare weighted score.

Step 6: Rank weighted score values to predict attributes.

Step 7: Evaluate the accuracy.

The first step is to transform the dataset into a new dataset, using the suggested approach, using the chosen optimal attributes for the decision making procedure, by retaining the details of the attribute the same. Next, the weight of the experts is obtained by entropy. The same ranking subset obtained is regarded as the better alternative for the separate values. The results are then calculated based on the current system of comparison.

The first is to combine the weighted value of the weighted score with the average weighted score operator. Secondly, for any alternate collective total data subset the score functions are estimated. Third, according to the score values, the best alternative is chosen. Finally it is possible to compute the score values and produce the ranking data. Various order positions are achieved with the various approaches, but the distinctions are subtle.

The findings of this work are also more rational, and the approach suggested is more flexible than the method proposed in the current study. The findings of the comparative study confirms that it is an accurate and realistic method to forecast medical diseases in the context of the decision-making phase in the healthcare system.

3.3 CLASSIFICATION

While CNN models have proven effective in many image processing scenarios, especially in medical image analysis, this has yet to be the case in object recognition. Inspired by this, we provide a CNN-based method for the identification and categorization of cervical cancer. To properly train CNN-based systems, massive amounts of medical image data are required. It is almost impossible to gather such a big library of medical pictures. Therefore, when the database size is small, it is more popular to use transfer learning and fine-tuning. To apply a deep CNN model, you must have a large amount of data to train it. Once the model is trained, you can utilise it as a pretrained model. To reach the required accuracy, the pretrained model parameters are fine-tuned using a selected database training set. This is a well-tuned model, which is used for testing.

With the suggested system, we perform three CNN training experiments, each of which consists of one of the models. The shallow architecture is involved in all three experiments, and the deep architecture is in two of the experiments.

There are only two convolutional layers and two max-pooling layers in the model. The RGB image, which is 224 pixels wide and 224 pixels high, is being used as the input. There are 64 different 5×5 filters in the first convolutional layer, and 128 different 5×5 filters in the second convolutional layer. The filter stride is 2 pixels. The mask size of the max-pooling filter is 2x2. The non-linear activation is carried out with the rectified linear unit. After the second max-pooling layer, the features are reduced to a single value (in contrast to many values), and they are fed into a fully linked layer. The bottom of the two fully linked layers and the top of the softmax (output) layer are connected. Once training was concluded, the fine-tuning of the database was completed using a separate training subset of the target database. A smaller subset of training data is utilised to train the model using Stochastic Gradient Descent (SGD) and optimise the model parameters. Batch size is 20, learning rate is 0.01, and epoch size is 50.

To understand the two deep CNN models in the form of the VGG-16 Net and the CaffeNet, we designed a new computer model that employs both of them. These two models were trained using a large number of photos and subsequently won many competitions. Additionally, they are commonly found in other applications, such as epilepsy seizure detection.

To supplement the ELM-based classifier, we have proposed using an EM classifier. The ELM is a shallow network, which offers several advantages, such as fast learning, simple convergence, and lower amount of randomizedness [49]. The system has two end-to-end load mimicking machines (ELMs) at the end of the last completely connected layer of the CNN model (removing the softmax layer). In the first ELM, the output is set to give normal or abnormal cells. In the second ELM, the output is set to give classifications of normal and abnormal cases. Once the first ELM training is finished, the output is discarded, and the hidden layer is used as the input to the second ELM. In order to calculate the dense representation, the number of neurons in the hidden layer is set to 2048, thus this ELM relates to a sparse representation. The SGD method is utilised to optimise the weight of the ELM.

4. RESULTS AND DISCUSSIONS

In the medical fields, several methods for calculating precision indexes that includes are used: sensitivity, specificity [34]. In this paper, the accuracy, sensitivity, and specificity for test data sets are calculated to assess the efficiency of classification model [15, and the variability thereof in several test sets obtained from randomized dataset partitions. The database was obtained from kaggle.com We used a 5-fold cross validation approach. For fine tuning, we used 80% of the data in iteration. The rest 20% was used for testing. After five iterations, all the data were tested. The final accuracy was obtained by averaging accuracies of five iterations. The Herlev dataset was used for both the training and the testing the shallow CNN model using the 5-fold cross-validation approach. There was no fine tuning the model.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad (2)$$

$$Sensitivity = \frac{TP}{TP+FN} \qquad (3)$$

$$Specificity = \frac{TN}{TN+FP} \qquad (4)$$

where *TP* is the number of cases correctly identified as patient, *FP* is the number of cases incorrectly identified as patient, *TN* is the number of cases correctly identified as healthy and *FN* is the number of cases incorrectly identified as healthy.

Table.1. Accuracy before applying feature selection Algorithm

Classifier	Accuracy	TP Rate	FP Rate	Precision	Recall
NB	91.17	0.93	0.08	0.72	0.93
J48	96.24	0.93	0.02	0.93	0.93
MLP	93.70	0.78	0.02	0.92	0.78
SVM	94.97	0.93	0.02	0.93	0.93

Table.2. ROC before applying feature selection Algorithm

Classifiers	F-measure	MCC	ROC Area
NB	0.810	0.775	0.948
J48	0.935	0.920	0.970
MLP	0.844	0.819	0.992
SVM	0.935	0.920	0.967

The feature selection allows us to improve the detection efficiency of the different classifiers with the appropriate model preparation, deleting all irrelevant and redundant features, and extracting the best possible features from the initial feature collection. In addition, we expected to reduce the number of original features and reduce the time required to train without compromise. The proposed weighted score significantly improved the cervical cancer prediction accuracy of different classifiers. NB, DT and ANN classifiers were significantly improved.

5. CONCLUSIONS

In this paper, ANN offers fast reactions by a quicker confirmation process. The ANN categorizes and forecast patient infection. In the analysis, the image data sets for classification are prepared using a preprocessing model and a feature extraction model. The simulation is carried out to test how the model works for image data sets of cervical cancer, i.e. human blood samples. The validation indicates that patients are efficiently classified more quickly than the other deep learning models by the proposed approach. In all performance measures, such as accuracy, accuracy, recall, and f-measure. ANN gives the best outcomes, the best classification classifier for a person with or without cervical cancer disease.

REFERENCES

[1] M. Guzman and G. Kouri, "Dengue and Dengue Hemorrhagic Fever in the Americas: Lessons and Challenges", *Journal of Clinical Virology*, Vol. 27, No. 1, pp. 1-13, 2003.

[2] S. Kannan and S.N. Mohanty, "Survey of Various Statistical Numerical and Machine Learning Ontological Models on Infectious Disease Ontology", *Proceedings of International Conference on Data Analytics in Bioinformatics: A Machine Learning Perspective*, pp. 431-442, 2021.

[3] N. Kousik, A. Kallam, R. Patan and A.H. Gandomi, "Improved Salient Object Detection using Hybrid Convolution Recurrent Neural Network", *Expert Systems with Applications*, Vol. 166, pp. 114064-114075, 2021.

[4] N.V. Kousik, "Analyses on Artificial Intelligence Framework to Detect Crime Pattern", *Proceedings of International Conference on Intelligent Data Analytics for Terror Threat Prediction: Architectures, Methodologies, Techniques and Applications*, 119-132, 2021.

[5] K. Srihari, S. Chandragandhi, G. Dhiman and A. Kaur, "Analysis of Protein-Ligand Interactions of SARS-Cov-2 Against Selective Drug using Deep Neural Networks", *Big Data Mining and Analytics*, Vol. 4, No. 2, pp. 76-83, 2021.

[6] K.M. Baalamurugan and S.V. Bhanu, "An Efficient Clustering Scheme for Cloud Computing Problems using Metaheuristic Algorithms", *Cluster Computing*, Vol. 22, No. 5, pp. 12917-12927, 2019.

[7] T. Karthikeyan and K. Praghash, "An Improved Task Allocation Scheme in Serverless Computing Using Gray Wolf Optimization (GWO) Based Reinforcement Learning

(RIL) Approach", *Wireless Personal Communications*, Vol. 80, No. 7, 1-19, 2020.

[8] J.L. San Martín, J.O. Solorzano and M.G. Guzman, "The Epidemiology of Dengue in the Americas over the Last Three Decades: A Worrisome Reality", *American Journal of Tropical Medicine and Hygiene*, Vol. 82, No. 1, pp. 128-135, 2010.

[9] K. Srihari, G. Dhiman, K. Somasundaram and M. Masud, "Nature-Inspired-Based Approach for Automated Cyberbullying Classification on Multimedia Social Networking", *Mathematical Problems in Engineering*, Vol. 2021, pp. 1-18, 2021.

[10] D.A. Thitiprayoonwongse, P.R. Suriyaphol and N.U. Soonthornphisaj, "Data Mining of Dengue Infection using Decision Tree", *Proceedings of International Conference on Latest Advances in Information Science and Applications*, pp. 1-14, 2012.

[11] V. Nandini and R. Sriranjitha, "Dengue Detection and Prediction System using Data Mining with Frequency Analysis", *Proceedings of International Conference on Computer Science and Information Technology*, pp. 1-12, 2016.

[12] G. Li, X. Zhou and J. Liu, "Comparison of Three Data Mining Models for Prediction of Advanced Schistosomiasis Prognosis in the Hubei Province", *PLoS Neglected Tropical Diseases*, Vol. 12, pp. 1-22, 2018.

[13] V. Chang, B. Gobinathan, A. Pinagapani and S. Kannan, "Automatic Detection of Cyberbullying using Multi-Feature Based Artificial Intelligence with Deep Decision Tree Classification", *Computers and Electrical Engineering*, Vol. 92, pp. 107186-107198, 2021.

[14] A. Shukla, G. Kalnoor and A. Kumar, "Improved Recognition Rate of Different Material Category using Convolutional Neural Networks", *Materials Today: Proceedings*, Vol. 78, No. 1, pp. 1-5, 2021.

[15] S. Kannan, G. Dhiman and M. Gheisari, "Ubiquitous Vehicular Ad-Hoc Network Computing using Deep Neural Network with IoT-Based Bat Agents for Traffic Management", *Electronics*, Vol. 7, No. 1, pp. 785-793, 2021.

[16] J. Gowrishankar, T. Narmadha and M. Ramkumar, "Convolutional Neural Network Classification On 2d Craniofacial Images", *International Journal of Grid and Distributed Computing*, Vol. 13, No. 1, pp. 1026-1032, 2020.

[17] A. Khadidos, A.O. Khadidos and G. Tsaramirsis, "Analysis of COVID-19 Infections on a CT Image using DeepSense Model", *Frontiers in Public Health*, Vol. 8, pp. 1-12, 2020.

[18] Siriyasatien Padet,Atchara Phumee, Phatsavee Ongruk, Katechan Jampachaisri and Kraisak Kesorn, "Analysis of Significant Factors for Dengue Fever Incidence Prediction", *BMC Bioinformatics*, Vol. 17, No. 166, pp. 1-22, 2016.

[19] P. Vivekanandan, "An Efficient SVM Based Tumor Classification with Symmetry Non-Negative Matrix Factorization using Gene Expression Data", *Proceedings of International Conference on Information Communication and Embedded Systems*, pp. 761-768, 2013.

[20] A. Daniel and K.M. Baalamurugan, "A Novel Approach to Minimize Classifier Computational Overheads in Big Data using Neural Networks", *Physical Communication*, Vol. 42, pp. 101130-101135, 2020.

[21] K.M. Baalamurugan and S.V. Bhanu, "A Multi-Objective Krill Herd Algorithm for Virtual Machine Placement in Cloud Computing", *The Journal of Supercomputing*, Vol. 76, No. 6, pp. 4525-4542, 2020.

[22] I. Kononenko, "Machine Learning for Medical Diagnosis: History, State of the Art and Perspective," *Artificial Intelligence in Medicine*, Vol. 23, No. 1, pp. 89-109, 2001.

[23] D. Raval, D. Bhatt, M.K. Kumar, V. Parikh and D. Vyas, "Medical Diagnosis System using Machine Learning", *International Journal of Computer Science and Communication*, Vol. 7, No. 1, pp. 177-182, 2016.

[24] M. Umar, D. Babu, K.M. Baalamurugan and P. Singh, "Automation of Energy Conservation for Nodes in Wireless Sensor Networks", *International Journal of Future Generation Communication and Networking*, Vol. 13, No. 3, pp. 1-12, 2020.

[25] C. Saravanabhavan, T. Saravanan, D.B. Mariappan, S. Nagaraj and K.M. Baalamurugan, "Data Mining Model for Chronic Kidney Risks Prediction Based on Using NB-CbH", *Proceedings of IEEE International Conference on Advance Computing and Innovative Technologies in Engineering*, pp. 1023-1026, 2021.