DeepLCP: Towards a DeepLearning Approach to Prevent Lung Cancer

Mayssa Ben Kahla, Dalel Kanzari, and Ahmed Maalel

1 Introduction

Lung cancer is considered to be one of the leading causes of death, mainly because of the late detection of the disease's symptoms and the lack of prevention's means. According to the National Cancer Institute: lung cancer is the fourth most common cancer in France. Also, according to the International Agency for Research on Cancer: Lung cancer is the second most common cancer in Tunisia with a 14.2% incidence and 21.1% mortality. This disease has many (a) risk factors, for example personal history of disease, family history of cancer, diet, smokers, etc., and many (b) symptoms, for example chest pain, persistent cough, Spitting blood, etc. The aim of our approach is to accurately calculate the probability of having lung cancer disease based on (a) and (b) by combining two technologies; the natural language processing (NLP) [1, 2] and the convolutional neural network (CNN) [3]. This paper describes in Sect. 2 the related work in Sect. 3 our approach, named DeepLCP, and in Sect. 4 the experimental validation of DeepLCP.

M. Ben Kahla (✉)
Higher Institute of Applied Science and Technology, University of Sousse, Sousse, Tunisia

D. Kanzari · A. Maalel
Higher Institute of Applied Sciences and Technology, University of Sousse, Sousse, Tunisia

National School of Computer Sciences, RIADI Laboratory, University of Manouba, Manouba, Tunisia

L. Chaari (ed.), *Digital Health in Focus of Predictive, Preventive and Personalised Medicine*, Advances in Predictive, Preventive and Personalised Medicine 12,
https://doi.org/10.1007/978-3-030-49815-3_3

2 Related Works

Several works deal with the cancer disease using the Deeplearning [4] paradigm, we quote, for example, the work of Gruetzemacher et al. [5] which use the architecture DNN for a revolutionary image recognition method to distinguish large and small pulmonary nodules from potentially malignant lung nodules.Besides uncertainty and high cost of computation, this work achieves high false positive in the detection step. Esteva et al. [6] deal with the CNN technique to classify the skin lesion and to detect the cancer disease by giving the probability of malignancy or benignity. But, this work complains by the variance of accuracy. Park et al. [7] develops a Deep learning algorithm, called DeepNEAT-Dx, to predict the presence or absence of lung cancer in a chest X-ray. The problem of "DeepNEAT-Dx", spend 40 h to train. Also, the study by D. Yang et al. [8], presents advanced AI (artificial intelligence) technology for the early detection of lung cancer and a classification model based on the DCNN system. Bychkov et al. [9] chose to combine convolutional (CNN) and recurrent (RNN) architectures to form a deep network to predict colorectal cancer outcomes from images of tumor tissue samples. But, this work has obtained low accuracy. All these works use image as input data, furthermore, there exist other works that use text as input data, such as those of Baker et al. [10] that apply the convolutional neural network (CNN) to classify the biomedical input texts. The disadvantages of this work the Cost of calculation and complexity. Also, John et al. [11] deal with the convolutional neural network (CNN), to extract ICDO-3 topographic codes from a corpus of breast and lung cancer pathology reports. The limitation of this corpus study included pathology reports for which the truth on the ground came only from the final diagnostic section of the report We summarize the different works that deal with deep learning, associated with text or image input data, to detect the cancer disease in Table 1.

The works presented in Table 1 contain a lot of limits like precision problem, cost calculation and complexity. Moreover, we note that most of these works apply convolutional neural networks (CNN) to detect only the lung cancer. They use CNN in a delayed phase where the patient made the diagnosis by imaging. For example, [7] used CNN to detect lung cancer from CXRs images and [11] use the CNN with medical reports after diagnosis. The problem with imaging, in case of lung cancer, is that the disease can't be discover early and the remedy is hardly difficult.

3 Proposed Approach

Inspired from Zhang Y. & Wallace (2015) [12] works, our approach, aim to combine two advanced methods; the natural language processing **(NLP)** and the convolutional neuronal network **(CNN)**. As illustrated in Fig. 1, our architecture is composed of an NLP layer, CNN layers, and a disease classification output.

Table 1 Synthesis of related works

Reference	Algorithm	Architecture	Accuracy	Error rate	Input image	Input text
[5]	Maxpooling, softmax	DNN	81.08–82.10%	_	✓	
[6]	Partitioning algorithm (PA)/Inference algorithm/t-SNE (t-distributed stochastic neighbor embedding)	CNN (GoogleNet)	55.4–72.1%	_	✓	
[7]	Genetic algorithm/ DeepNEAT-Dx/ Backpropaga-tion/SchiffMan encoding	CNN	96.00%	7.97%	✓	
[8]	_	DCNN	90–97%	10–24%	✓	
[9]	SVM/Naïve Bayes classifier/logistic regression	CNN-RNN (LSTM)	69%	_	✓	
[10]	NLP-WORD2VEC word embeddings	CNN	97.1%	2.9%		✓
[11]	Adadelta adaptive gradient descent/NLP-WORD2VEC/N-grams	CNN	71%	30%		✓

3.1 Natural Language Processing (NLP)

In the **NLP** part we use the word2vec model [13] to convert the sentences extracted from our online form to raw matrix. each sentence is a feature. The weights will choose according to the user's response and semantic transformation rules.

Semantic Transformation Rules give each information weights. All the rules defined by two doctors from hospital Farhat Hached Sousse and Hospital Taher Sfar Mahdia. We use 31 semantic rules formed by the formal Z language[14], then we implement these rules in python the construction of our raw semantic matrix. Figure 2 shows an example of these rules.

This rule means if the person answered in our online form that his gender is a man then the first column weight of VC matrix is a value of the interval [0.6..0.9]. Else if the answer is a woman then the weight of the first VC matrix column is a value

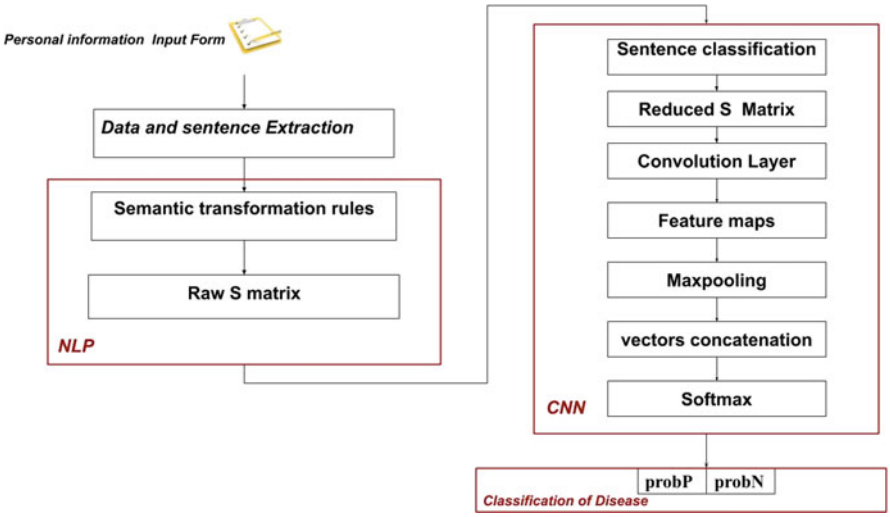

Fig. 1 DeepLCP architecture

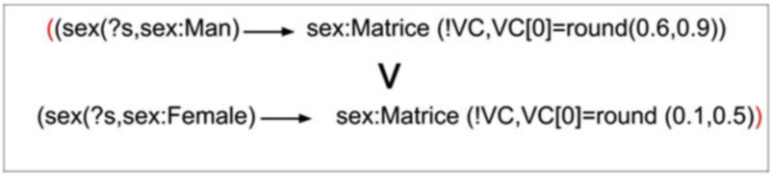

Fig. 2 Semantic transformation rules

of the interval [0.1..0.5]. These intervals are suggested by the doctor "Pr. Bouaouina Noureddine" chief radiotherapy department in the hospital Farhat hached Sousse and the doctor "Dr. Jalel Knani" Pneumologist in the Tahar Sfar Hospital because the Man have the risk of having the disease that the woman.

Raw S Matrix after the transformation we obtain the raw semantic matrix as illustrated in Fig. 3 with size [31*13].

- **31**: is the number of features.
- **13**: is the maximum number of words in the longest sentence.

3.2 Convolutional Neural Network (CNN)

In this part we apply two semantic classifications on the raw semantic matrix to obtain the reduced semantic matrix. Then we apply the CNN on this matrix to obtain the probability of detecting the disease.

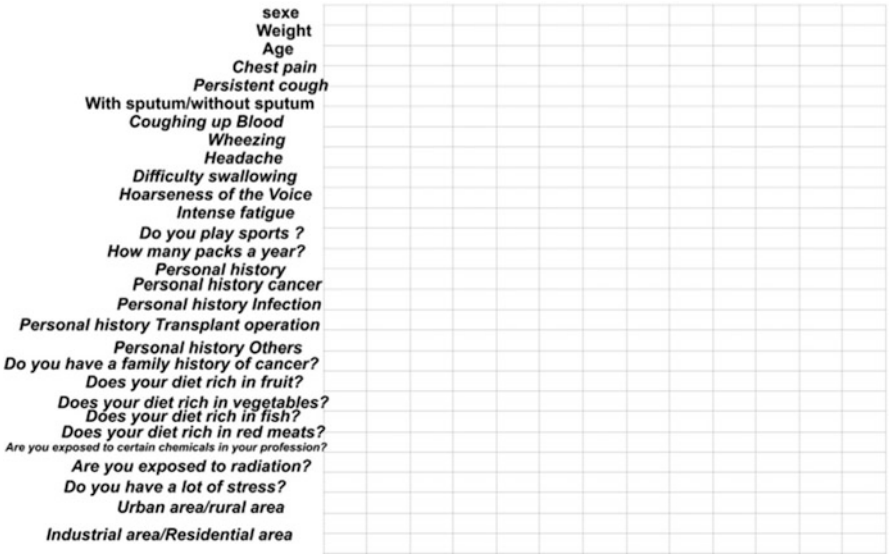

Fig. 3 Raw semantic matrix

- **Reduced S Matrix**
  To obtain the reduced semantic matrix we apply two type of classifications:
  - **Classification By Categories:** we classify data according to three categories: **Minor risk factors, Major risk factors, symptom**.
  - \+ **Classification By Themes:** we classify data into six themes:
    - **Thoracic signs:** this matrix represents the average of the matrices "chest pain", "wheezing" and "abnormal breathlessness".
    - **Cough:** this matrix contains the average of the matrices of "persistent cough", "with/without spitting" and "spitting of blood".
    - **Feeding:** this matrix contains the average of the matrices "diet rich in fruits", "diet rich in vegetables", "food rich in fish" and "diet rich in red meat".
    - **Consumer:** this matrix contains the average of the matrices "how many packets tobacco", "passive smoking", "alcohol".
    - **Personal antecedent:** this matrix contains the average of the matrices "Cancer", "infection", "Transplant operation of an organ".
    - **Residence:** this matrix indicates either the average of the matrices "urban area" and "industrial zone", either the average of the matrices "urban area" and "residential area", either the average of the matrices "rural area" and "residential area".

After classification we obtain the reduced semantic matrix with size [18*13] as illustrated in Fig. 4.

Fig. 4 Reduced matrix

We used the reduction to optimize the complexity. With the reduction will not lose any information.

- **Convolution Layer**
  In this layer we have 3 region size (5, 6, 7) and 2 filters for each region size. So, we have six filters in the total. All the filters will scan all the matrix input with stride 1 to give us one feature maps for each filter. For each feature map we will apply the maxpooling to get one value for each feature map accordingly concatenate all the maxpooling and apply the activation function "softmax".
- **Disease Classification** Our output is two probability "probP": probability of having the disease and "probN": probability of not having the disease.

4 Validation

In the validation part, we use validation set with 490 real cases divided into 315 of the patient who have the disease and 175 who haven't. Also, we use 111 real cases for test set divided into 40 patients affected by the disease and 71 who not affected. As it is shown in Figs. 5 and 6 the accuracy for our model is 94.59% with a 5.41% error rate and the test lost 15.90%.

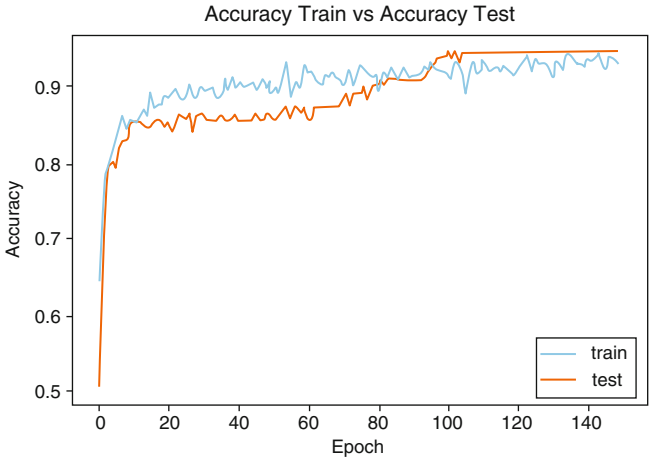

Fig. 5 Accuracy train vs accuracy test DeepLCP

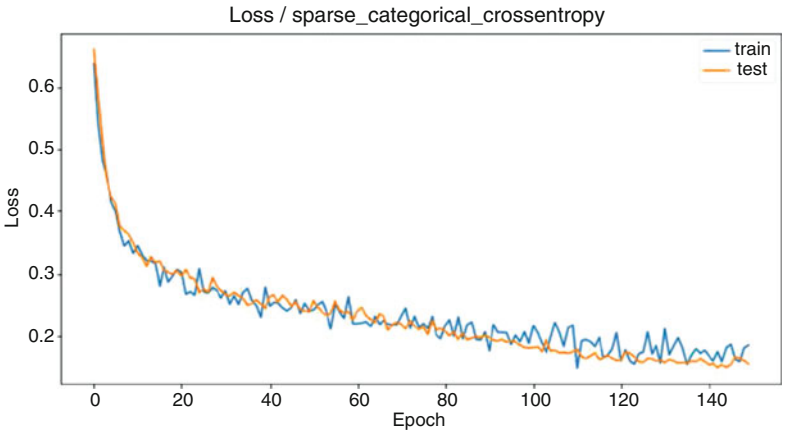

Fig. 6 Loss train vs loss test DeepLCP

5 Discussion

We tested our dataset with four machine learning algorithms:

- The **k-nearest neighbors (KNN)**: it provides a 86.48% precision rate and a 13.52% Error rate.
- The **Decision Tree algorithm**: it returns a 93.69% precision rate and a 6.31% error rate.
- The **Random Forest**: it results a 91.89% precision rate and a 8.11% error rate.
- The **Artificial Neural Network (ANN)**: it provides a 85.59% precision rate and a 14.41% error rate.

Based on these results we find that our "DeepLCP" model provides the best accuracy rate and the lowest error rate.

6 Conclusion

In this article we present a new model for the prevention of lung cancer. Our model named "DeepLCP" is a combination of NLP and CNN. In the NLP part we use semantic transformation rules. the accuracy of validation test is 94.5% which confirm that our model give an efficient result.

References

1. Otter, D.W., et al.: A survey of the usages of deep learning in natural language processing. CoRR, abs/1807.10854 (2018)
2. Towards Datascience: https://towardsdatascience.com/natural-language-processing-nlp-for-machine-learning-d44498845d5b. Last accessed 13 June 2019
3. Analytics Vidhya: https://www.analyticsvidhya.com/blog/2018/12/guide-convolutional-neural-network-cnn/. Last accessed 14 June 2019
4. Pattanayak, S.: Pro Deep Learning with TensorFlow: A Mathematical Approach to Advanced Artificial Intelligence in Python. Apress, Berkeley (2018)
5. Gruetzemacher, R., Gupta, A.: Using deep learning for pulmonary nodule detection & diagnosis. In: AMCIS, Association for Information Systems, San Diego (2016)
6. Esteva, A., et al.: Dermatologist-level classification of skin cancer with deep neural networks. Nature **542**, 115–118 (2017)
7. Michael Park, H., Monahan, C.: Genetic deep learning for lung cancer screening. Innovation Dx Inc. 23 Aug (2017)
8. Yang, D., Powell, C.A., et al.: Deep convolutional neutral networks based artificial intelligence system for pulmonary nodule detection and diagnosis in United States and Chinese dataset. In: ATS, San Diego (2018)
9. Bychkov, D., et al.: Deep learning based tissue analysis predicts outcome in colorectal cancer. In: Scientific Reports, Feb (2018)
10. Baker, S., et al.: Cancer hallmark text classification using convolutional neural networks. In: BioTxtM@COLING (2016)
11. John, X., et al.: Deep learning for automated extraction of primary sites from cancer pathology reports. IEEE J. Biomed. Health Inf. **22**, 244–251 (2018)
12. Zhang, Y., Wallace, A.: Sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. arXiv preprint arXiv:1510.03820 (2015)
13. Dataanalytics Post: https://dataanalyticspost.com/Lexique/word2vec/. Last accessed 14 June 2019
14. Bowen, J.P.: Formal Specification and Documentation Using Z: A Case Study Approach. International Thomson Publishing, London/Boston, June (2003)