Adaptive Information Assimilation using Convolutional Neural Network for Forecast of Breast Cancer from Electronic Health Records

Senthil Kumar Kumar J (senthilkumarj@mepcoeng.ac.in)
Mepco Schlenk Engineering College https://orcid.org/0000-0002-9516-0327
Kamala Devi K
Mepco Schlenk Engineering College
Raja Sekar J
Mepco Schlenk Engineering College

Research article

Keywords: Breast Cancer, Pathology Reports, Deep learning, Convolutional neural network

Posted Date: December 2nd, 2019

DOI: https://doi.org/10.21203/rs.2.18006/v1

Abstract

Purpose Data acquired from cancer based Electronic Health Records (EHRs) shows key statistics on cancer affected persons. To estimate the impact of the cancer on those persons, we need to extract vital information from those pathology health records. It is an exhaustive procedure to carry out because of large volume of records and data acquired for a continuous period of time.Methods This research portrays, the investigation of convolutional neural network (CNN) and Support Vector Machine (SVM) techniques for extracting topographic codes from the pathology reports of breast cancer. Investigations are carried out using conventional frequency vector space method and the deep learning techniques such as CNN. The learning experience of those algorithms were absorbed on a set of 730 pathology reports.Results We perceived that the CNN technique reliably outperformed the conventional frequency vector methods. It is also observed that it causes the micro and macro average performance to increase up to 0.119, and 0.101, while considering the populated class labels for the CNN model. Unambiguously, the top performing CNN approach attained a micro-F score of 0.821 over the considered topography codes.Conclusion These promising outcomes reveals the prospective of deep learning approaches, particularly CNN for estimating the impact of the cancer from the pathology reports compared to conventional SVM approach. More advanced and accurate approaches to effectively improve the accuracy in information extraction are needed.

Background

Health care organizations have started using the Information Technology service in our contemporary society. It has become a mandate choice for many clinical and administrative activities. Usage of Electronic Health Records (EHR) have started playing a significant role for such tasks. For extracting valuable information from EHR data deep learning techniques can be applied (Benjamin et al. 2017). EHRs are also used for making decisions about affected tissues after thorough examinations of them. Decisions from EHRs, are vital for the patient's present and feature health issues. Pathologist use invasive methods on the patients as one technique for obtaining biopsy from affected tissues of human body. They can also do the review by sneaking through the pathology records in EHRs. Primary view of the symptoms from those EHRs can make the pathologist to take appropriate decisions and give proper directions and medications to the patients. Extracting valuable information on the disease form the EHRs is quite challenging, if the volume of data in the EHRs is larger. Lot of researches have been carried out to manage those data and extract the information for performing accurate clinical decisions by pathologists.

For more accurate prediction of health relevant parameters from human body, large volume of data need to be analyzed. Big data technologies can support the health care industries by processing a large volume of EHR data for extracting vital health parameters from the patients (Marco et al. 2015). It enables to estimate sensitive information that cannot be easily determined with individual patient data. EHR holding the physiological variables for patients of different age group and health conditions are smart enough for analyzing and predicting the diseases. Researchers mostly find it challenging to collect

detailed information about patients. So, the publically available SEER cancer data is mostly used for training the developed models and estimating accurate insights from them. It includes health record of almost 28% of population in US (SEER, 2016). Raw data from SEER based EHR cannot be directly utilized for processing. The dataset is obtained by proper signed agreement from the National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) Program.

Wearable devices with low power consumption are gaining popularity for healthcare applications. Using spectral preprocessing on the health data from the wearable devices, it makes the data ready for processing by deep learning frameworks (Daniele et al. 2017). Advent of Internet of Things (IoT) gave made a significant impact for the wide spread usage of wearable health care devices. Data acquired from those smart wearable health care devices also support to large extent to generate health records.

Deep Neural Networks (DNN) are gaining popular the support of deep learning techniques. Those networks have multiple layers, which are capable of extracting meaningful features and learning from the data. Fig.1 shows the structure of a simple DNN, which has an input and output layers along with multiple hidden layers. Learning occurs with the support of forward and backward propagation of the weights associated with each neuron. Deep learning approaches using the EEG data are employed for reducing sleeping disorders of human. From its results diagnosis of insomnia can be done effectively (Mostafa et al. 2017).

Convolutional Neural Networks are a category of DNN. It has been gaining popular with the usage of medical images for classification and prediction of diseases. It can also be used for extracting deep hierarchical features in medical images. Convolutional Neural Networks are used for cervical cancer screening (Ling et al. 2017), Breast Cancer screening (Moi et al. 2017), for diagnosis of auto immune diseases (Zhimin et al. 2017), risk prediction using EMR (Phuoc et al. 2017), Gait parameter extraction (Julius et al. 2017), Lung tissue classifications (Qiangchang et al. 2018), and extracting primary cites from pathology reports (John et al. 2018) and many more. Fig.2 shows the general Architecture of a simple CNN considering the breast pathology report as input document matrix. The convolution layer observers the features from the pathology reports, the pooling layer eliminates the redundant features and the prediction of the tumor is performed at the final softmax output stage of the network.

Recurrent Neural Networks (Edward et al. 2016), are used for assisting doctors by providing best clinical decision support using EHR. From the large volume of patient records RNN based models predicts the future impacts of the diseases to the patients (Trang et al. 2017). Deep dynamic neural networks are also employed to predict future consequences of patients health from their health records. Long short team memory (LSTM) uses memory of historical records with time stamps for predicting the future risk factors (Truyen et al. 2015).

Medical record data objects are embedded using low dimensional vector space using Restricted Boltzmann Machine (RBM). This mechanism is encouraged for usage in medical records that are mostly

discrete in nature. By offering embedded space, most knowledge on the data can be exploited. Hierarchical order of learning from medical data are also carried out, by considering the order of visit and co-occurrence of the medication codes during the visits to health care organizations. It is made viable by considering an architecture with multilayer perceptron (MLP) (Edward et al. 2016).
Breast cancer survival rates of both women and men were analyzed from health records (Paulo et al. 2017). The analysis were performed using descriptive statistics on Cox regression and Kaplan-Meier analysis. Apart from gender, their drinking, smoking and other habits are also observed to predict the overall survival and disease free survival of the patients.

Early stages of recurrences in breast cancer are analyzed for estimating the risks and follow up actions for the cancer detected cases (Vinzenz et al. 2019). From the health data collected from German and Dutch people, analysis are performed by considering biological subtypes, surgery type and age of the patients. Decision making during breast cancer treatment need to consider the biological subtype and the pattern of recurrence at different time intervals (Ignatov et al. 2018). With known biological subtype, analysis were performed with the health care data. It is observed that the tumours that were initially low, remains the same even after 10 years. Observation from the enriched biological subtypes shows the increased pattern tumour.

Precision in cancer prediction can be improved by using an appropriate production model implemented using machine learning techniques. Penalized regression technique is implemented to build a predictive model to observe the resistance of epidermal growth factor receptor tyrosine kinase inhibitors (Young et al. 2018).

The research article starts with presenting a brief introduction to usage of EHR data, and deployment of big data and deep learning techniques for modern health care applications. Section II deals with the health data set and the strategies made for information extraction from EHR. Section III discusses the methodology used for training and testing the cancer pathology reports extracted from SEER dataset using CNN. Section IV summaries the results analyzed after the successful deployment of the CNN for effective information extraction from the health records. Finally, Section V summarizes the work in the conclusion part.

1. HEALTH DATASET FOR INFORMATION ASSIMILATION

The developed multi study derived model for prediction provides good transferability and generalizability along with perfect accuracy during observations. Incidence of brain metastases is observed from the SEER datasets for production of prognosis (Yi-Jun et al. 2018). From a large set of breast cancer patients details collected from the SEER data, more incidence of brain metastases is observed for HER_2 subtype. Also visceral metastases is observed from the patients having TNBC and HER_2 subtypes. This analysis contributes to earlier metastases and positively increases the survival rate of the affected breast cancer patients.

From the SEER dataset, patients with stage IV breast cancer were identified and they the clinical value of auxiliary lymph nodes were assessed (San-Gang et al. 2017). The effect of the auxiliary lymph node dissection with the survival rates of the patients were analysed. It was observed that the auxiliary lymph node dissection improves with the survival rate of the patients.

From the health records, it is perceived that dissection of axillary lymph node improves the survival rate of patients diagnosed with breast cancer of stage IV. This observation was made on the patients who received tumor surgery in primary case, especially in liver and bone (Wu SG et al. 2017). Gender based survival of breast cancer analysis was performed with the hospital health records. Overall survival and disease free survival studies on them reports no noteworthy dissimilarity in prediction, but changes in clinical features were founded based on their demographic locations (Thuler et al. 2017). With familiar biological subtype of the patients, the breast cancer recurrence patterns are studied. It is evident that with varying time their subtypes are changed accordingly and they need to be considered while making a decision about tumour (Ignatov et al. 2018). From the SEER dataset prediction of brain metastasis is performed from the breast cancer reports. It is evaluated based on the molecular subtypes and estimated that patients with TNBC and HER_2 subtypes possess visceral metastasis (Kim et al. 2018)

Pathology reports from SEER dataset that matches with 730 cases of breast cancer are chosen or analysis from the registry. The topography codes used in the training set of the analysis process includes only the final diagnosis part from the pathology report. This kind of choice is made to avoid variation during the training process and to improve the robustness of the estimations from the reports. Table 1 shows the 9 ICD-O-3 topography codes that includes the primary sites of breast cancer chosen for the analysis. In the preprocessing stage, the text contents of the pathology report are aggregated to carefully utilize the empty sections in the reports.

Methods

Extraction of valuable information from the SEER health records can be performed by fragmenting the sequence of EHR data and by performing Multi hot encoding of the sequence. Fig.3 shows the sequence of steps to be carried out for feeding the extracted data from EHR for performing the processing using deep neural networks.

Corpus of data can be encoded using feature vectors based on the count of words. This vector space models are basic tasks of NLP systems for relatively simpler extraction of vital health care information from the data set. Based on the observation similarities, the word embedding techniques can be used for information extraction.

Usage of deep learning techniques to learn the representation of words from the data set, unlike conventional observation methods can provide better accuracy and minimizes the efforts in information retrieval.

Few earlier works on extracting of text data using deep learning techniques uses recurrent networks (Mikolov.T, et. al. 2010). Some of the literatures also extracted the data using feature vectors on the encoded documents (Le. Q. et. al. 2014). These category of information extraction largely depends on the structure and form of the documents used. Even though CNN were developed for vision based tasks in deep learning approaches, it has found its deep rooted impact on NLP, and literatures have utilized its extraordinary performance for information extraction from documents (Zhang. Et. al. 2016). Also, utilization of the convolution filters in CNN and its max pooling techniques for information extraction from documents improves the accuracy when compared to the conventional techniques. It is highly applicable for features with higher dimensions and it can utilize the order of words in the document directly.

In the proposed investigation, we use the word segmentation process and word vector representation to the train the classifier using deep learning technique. The process for training and sequence extraction for tumor prediction is illustrated in the Fig. 4.

The sequence of word vectors is trained to maximize the objective function for a word of context. Trained vector of words after the word segmentation are able to capture different meanings of the words in the context.

Analysis And Results

Analysis of the extracted pathology reports from SEER database are performed to test the effectiveness of the proposed DNN. In this research paper, we study the effectiveness of our proposed framework on SEER EHR data. From the extracted SEER dataset F1-score, precision and accuracy, were used to estimate the efficiency and performance of the proposed CNN framework architecture. For estimating much better performance, recall and precision measures are joint together obtain a better thoughtful understanding of the classifier. They are computed using the following expressions shown from eq(1) to eq(5).

$$Accuracy = \frac{tn+tp}{fn+tn+fp+tp} \quad (1)$$

$$Precision = \frac{tp}{fp+tp} \quad (2)$$

$$Recall = \frac{tp}{fn+tp} \quad (3)$$

$$F1-score = \frac{2 \times Precision \times Recall}{recision \times Recall} \quad (4)$$

$$Specificity = \frac{tn}{fn+tn} \quad (5)$$

Were is true negative, representing total predicted affected region, is true positive, is false negative and is false positive. Accuracy defined in eq(1) depicts the classification success rate considering both true and false values. The precision shown in eq(2), computes only with respect to the positive outcomes of the classifier. Similarly, the eq(5), computes only with respect to the negative outcomes. Performance

evaluation measure are dominant while considering the F1-score of the classifier. Both micro and macro F1-score are computed from the eq(4).

The proposed CNN architecture follows the implementation stages as shown in Fig.5. Initially the random weights are initialized and combined with the patient pathology records. The features are being associated to the input nodes. Followed by this initial stages, forward propagation is carried out by calculating the error function and predicting the design. This process is repeated to activate the neurons in the network based on the updated weights with respect to the error, till reaching the desired result.

Updating of the weights are performed in the backpropagation stage with the support of the backpropagation ID. The weights are repeated to update each data input, till reaching the desired result. The entire task is repeated for the training set and the process is done for multiple epochs. Once the desired accuracy is reached the process is stopped.

Quality of the classifications are evaluated using the confusion matrix. Accuracy in the predictions are observed from the diagonal values of the matrix. Normalized confusion matrices are plotted for SVM and CNN based observations. Analysis are also performed for minimally populated tasks and well-populated tasks.

From the normalized confusion matrix shown in Fig.6 and Fig.7, it is evident that the proposed CNN based classifier outperforms the SVM classifier to a larger extent for the minimally populated tasks. The diagonal elements of the figures represent the true positive classification performed successfully. The vertical elements specify the false positive classification performed. The false negatives are represented in the horizontal axis of the confusion matrix. It is evident that the CNN model classifies better for the breast classes c34.0, c34.1 and c34.9 in the minimally populated classes

Similar form of normalized confusion matrix shown in Fig.8 and Fig.9, is plotted for the well populated tasks. From the observations, it is evident that the proposed CNN based classifier outperforms the SVM classifier with better accuracy. The diagonal elements of the CNN confusion matrix for well populated tasks shows more true positives compared to SVM technique. It is evident that the CNN model classifies better for the breast classes c34.1, c34.3 and c34.9 in the well populated classes.

Table 2 shows the comparison and consolidated results of the SVM and CNN based classifiers of the different pathology reports with breast cancer. It is shown for both minimally and well populated tasks. From the eq(1) to eq(5), the performance measures are calculated and they are tabulated with accuracy, precision, F-score and specificity. From the observations made, the CNN model outperforms the SVM for breast cancer information assimilation for both minimally and well populated tasks.

This kind of classifiers are better choice for information extraction from the electronic health record. Deep learning based CNN model show cases with appropriate well defined strategy with better accuracy than the conventional SVM classifier.

Conclusions

In this proposed research article, we have designed and developed a deep neural network for information extraction from pathology reports. Series of experiments were done using CNN and traditional SVM classifiers on the SEER dataset. The performance of CNN is observed to be superior with better micro-F and macro-F scores of 0.821 and 0.794 respectively. Assimilation of information from the highly populated class of embedded randomized data in the CNN layers leads to better performance than the SVM classifiers.

Declarations

The authors would like to thank the Department of Computer Science and Engineering, and the Management, Principal of Mepco Schlenk Engineering College, Sivakasi, Tamil Nadu, India for providing us the modern state-of-art facilities to carry out this research work.

Compliance with ethical standards

Conflict of interest: The authors confirm that they have no conflict of interest regarding this research article

References

Shickel B, Tighe PJ, Bihorac A, Rashidi P. Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. *IEEE journal of biomedical and health informatics.* 2017;22(5):1589-1604.

Viceconti M, Hunter P, Hose R. Big data, big knowledge: big data for personalized healthcare. *IEEE journal of biomedical and health informatics.* 2015;19(4):1209-1215.

Ravi D, Wong C, Lo B, Yang G-Z. A deep learning approach to on-node sensor data analytics for mobile or wearable devices. *IEEE journal of biomedical and health informatics.* 2016;21(1):56-64.

Shahin M, Ahmed B, Hamida ST-B, Mulaffer FL, Glos M, Penzel T. Deep learning and insomnia: assisting clinicians with their diagnosis. *IEEE journal of biomedical and health informatics.* 2017;21(6):1546-1553.

Zhang L, Lu L, Nogues I, Summers RM, Liu S, Yao J. DeepPap: deep convolutional networks for cervical cell classification. *IEEE journal of biomedical and health informatics.* 2017;21(6):1633-1643.

Yap MH, Pons G, Martí J, et al. Automated breast ultrasound lesions detection using convolutional neural networks. *IEEE journal of biomedical and health informatics.* 2017;22(4):1218-1226.

Gao Z, Wang L, Zhou L, Zhang J. HEp-2 cell image classification with deep convolutional neural networks. *IEEE journal of biomedical and health informatics.* 2016;21(2):416-428.

Nguyen P, Tran T, Wickramasinghe N, Venkatesh S. $\mathtt {Deepr} $: a convolutional net for medical records. *IEEE journal of biomedical and health informatics.* 2016;21(1):22-30.

Hannink J, Kautz T, Pasluosta CF, Gaßmann K-G, Klucken J, Eskofier BM. Sensor-based gait parameter extraction with deep convolutional neural networks. *IEEE journal of biomedical and health informatics.* 2016;21(1):85-93.

Wang Q, Zheng Y, Yang G, Jin W, Chen X, Yin Y. Multiscale rotation-invariant convolutional neural networks for lung texture classification. *IEEE journal of biomedical and health informatics.* 2017;22(1):184-195.

Qiu JX, Yoon H-J, Fearn PA, Tourassi GD. Deep learning for automated extraction of primary sites from cancer pathology reports. *IEEE journal of biomedical and health informatics.* 2017;22(1):244-251.

Choi E, Bahadori MT, Schuetz A, Stewart WF, Sun J. Doctor ai: Predicting clinical events via recurrent neural networks. Paper presented at: Machine Learning for Healthcare Conference2016.

Pham T, Tran T, Phung D, Venkatesh S. Deepcare: A deep dynamic memory model for predictive medicine. Paper presented at: Pacific-Asia Conference on Knowledge Discovery and Data Mining2016.

Tran T, Nguyen TD, Phung D, Venkatesh S. Learning vector representation of medical objects via EMR-driven nonnegative restricted Boltzmann machines (eNRBM). *Journal of biomedical informatics.* 2015;54:96-105.

Choi E, Bahadori MT, Searles E, et al. Multi-layer representation learning for medical concepts. Paper presented at: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining2016.

Wu S-G, Li F-Y, Chen Y, et al. Therapeutic role of axillary lymph node dissection in patients with stage IV breast cancer: a population-based analysis. *Journal of cancer research and clinical oncology.* 2017;143(3):467-474.

Bender PFM, de Oliveira LL, Costa CR, de Aguiar SS, Bergmann A, Thuler LCS. Men and women show similar survival rates after breast cancer. *Journal of cancer research and clinical oncology.* 2017;143(4):563-571.

Voelkel V, Draeger T, Groothuis-Oudshoorn CG, et al. Predicting the risk of locoregional recurrence after early breast cancer: an external validation of the Dutch INFLUENCE-nomogram with clinical cancer registry data from Germany. *Journal of cancer research and clinical oncology.* 2019:1-11.

Ignatov A, Eggemann H, Burger E, Ignatov T. Patterns of breast cancer relapse in accordance to biological subtype. *Journal of cancer research and clinical oncology.* 2018;144(7):1347-1355.

Kim YR, Kim SY. Machine learning identifies a core gene set predictive of acquired resistance to EGFR tyrosine kinase inhibitor. *Journal of cancer research and clinical oncology.* 2018;144(8):1435-1444.

Gordon-Dseagu VL, Devesa SS, Goggins M, Stolzenberg-Solomon R. Pancreatic cancer incidence trends: evidence from the Surveillance, Epidemiology and End Results (SEER) population-based data. *International journal of epidemiology.* 2017;47(2):427-439.

Hill F, Cho K, Korhonen A. Learning distributed representations of sentences from unlabelled data. *arXiv preprint arXiv:160203483.* 2016.

Kim Y-J, Kim J-S, Kim IA. Molecular subtype predicts incidence and prognosis of brain metastasis from breast cancer in SEER database. *Journal of cancer research and clinical oncology.* 2018;144(9):1803-1816.

Lee H-Y, Tseng B-H, Wen T-H, Tsao Y. Personalizing recurrent-neural-network-based language model by social network. *IEEE/ACM Transactions on Audio, Speech, and Language Processing.* 2016;25(3):519-530.

Zhang D, Wang J, Zhao X. Estimating the uncertainty of average F1 scores. 2015.

Zhang Y, Wallace B. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. *arXiv preprint arXiv:151003820.* 2015.

Tables

Table 1: Breast ICD-O-3 Topographical Codes with Description and Count

Breast ICD-O-3 Topographical Codes	Description	Count
C50.0	Nipple	53
C50.1	Central Portion of breast	61
C50.2	Upper inner quadrant of breast	172
C50.3	Lower inner quadrant of breast	46
C50.4	Upper outer quadrant of breast	108
C50.5	Lower outer quadrant of breast	39
C50.6	Axillary tail of breast	39
C50.8	Overlapping lesion of breast	165
C50.9	Breast NOS	47

Table 2: Comparison of the SVM with proposed CNN model for minimally and well populated tasks.

Performance Indices	SVM		Proposed CNN algorithm	
	Minimally Populated Task	Well Populated Task	Minimally Populated Task	Well Populated Task
Accuracy (%)	67.50	71.67	75.41	85.00
Precision/ Sensitivity (%)	67.84	72.14	76.81	85.67
Micro F-Score	0.669	0.702	0.721	0.821
Macro F-Score	0.605	0.693	0.712	0.794
Specificity (%)	68.64	72.87	77.58	86.95

Figures

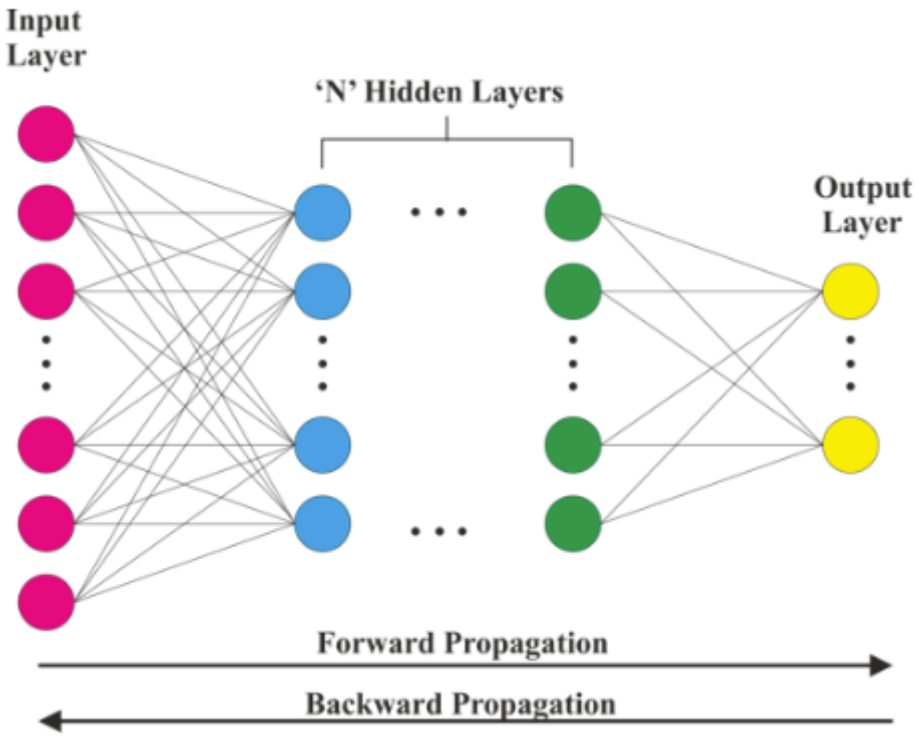

Figure 1

Structure of a Simple Deep Neural Network.

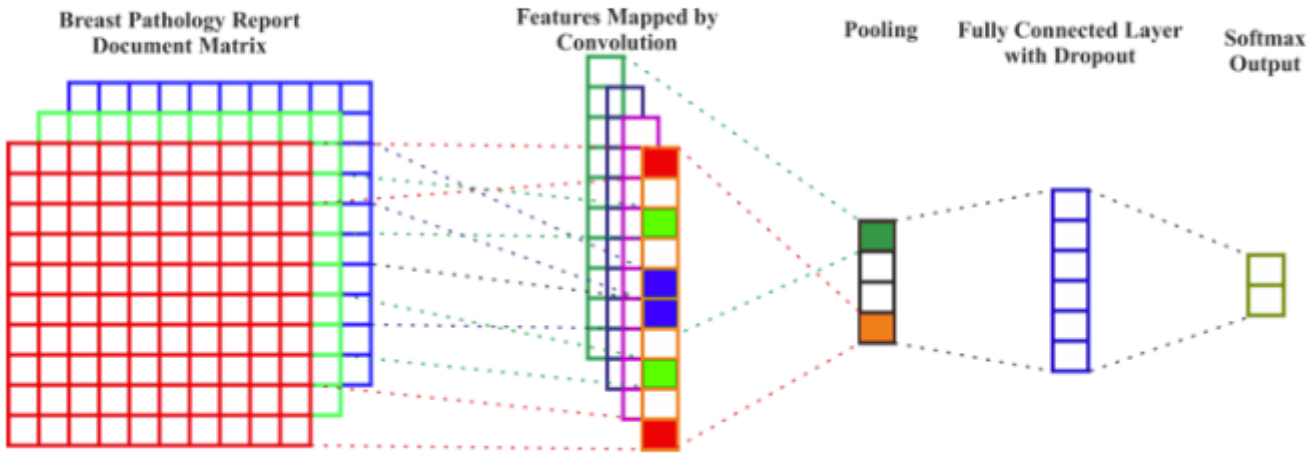

Figure 2

Architecture of Convolutional Neural Network

Figure 3

Data Extraction sequence from the SEER EHR Data set

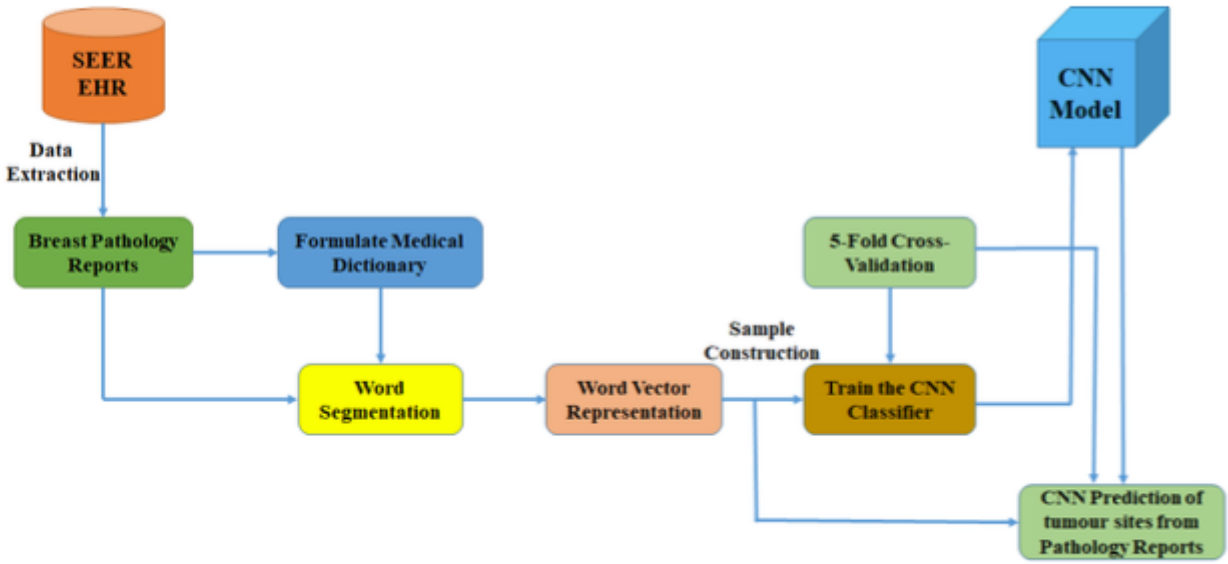

Figure 4

Training and Testing Sequence of Extracted Pathology reports for prediction by CNN

Figure 5

Stages in the proposed CNN Deep Neural Network e of Extracted Pathology reports for prediction by CNN

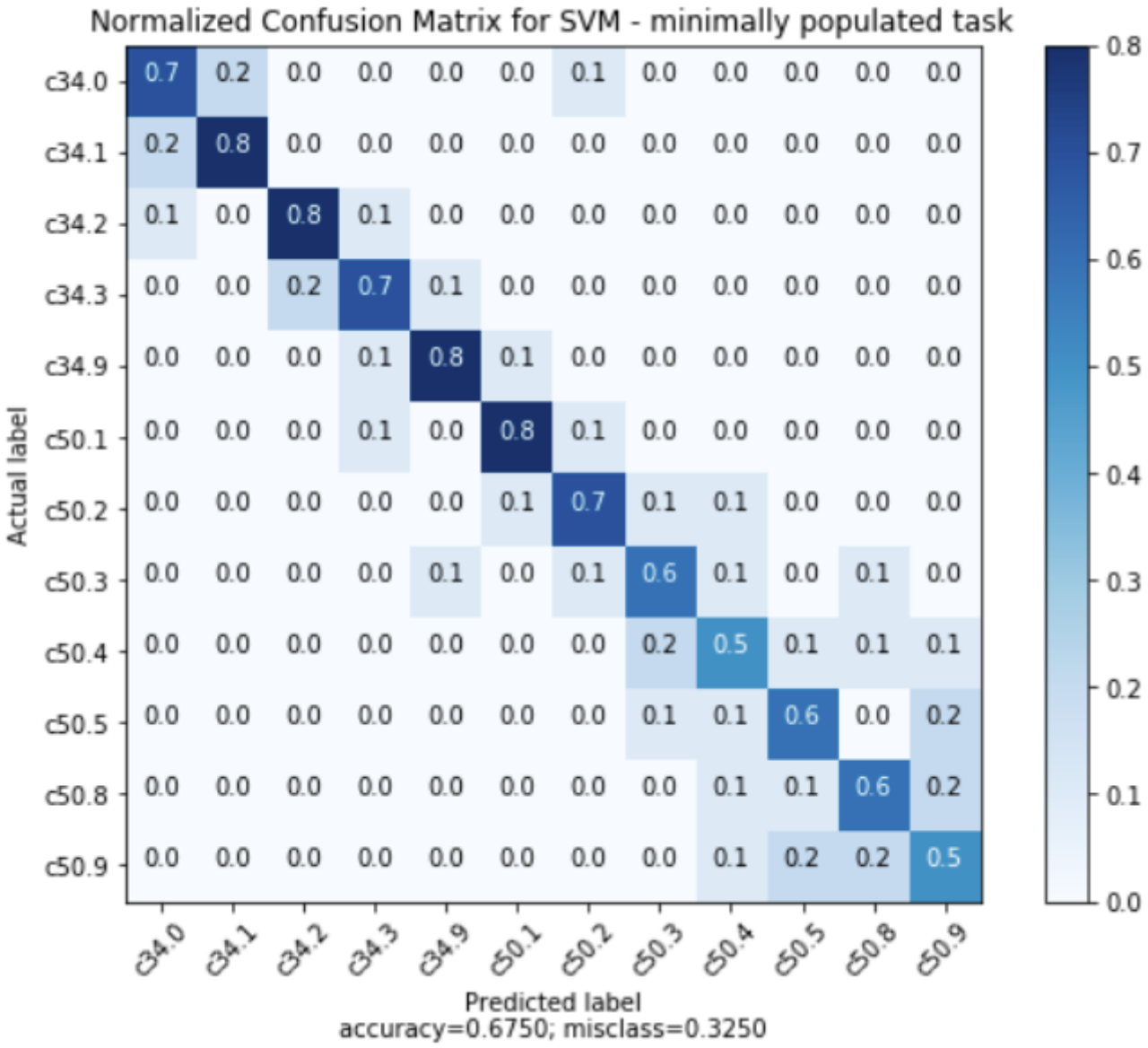

Figure 6

Normalized Confusion Matrix for SVM - minimally populated task

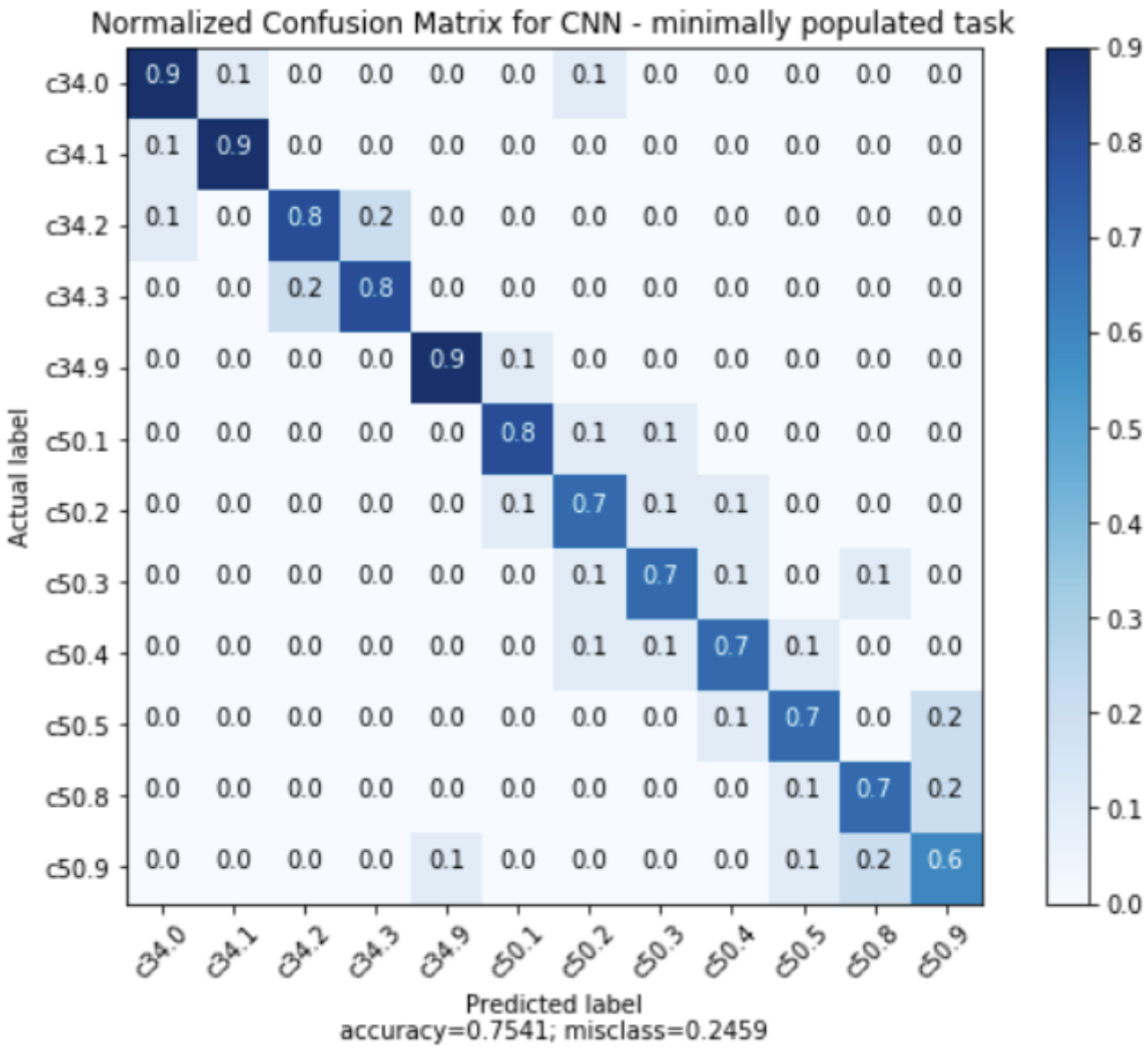

Figure 7

Normalized Confusion Matrix for CNN - minimally populated task

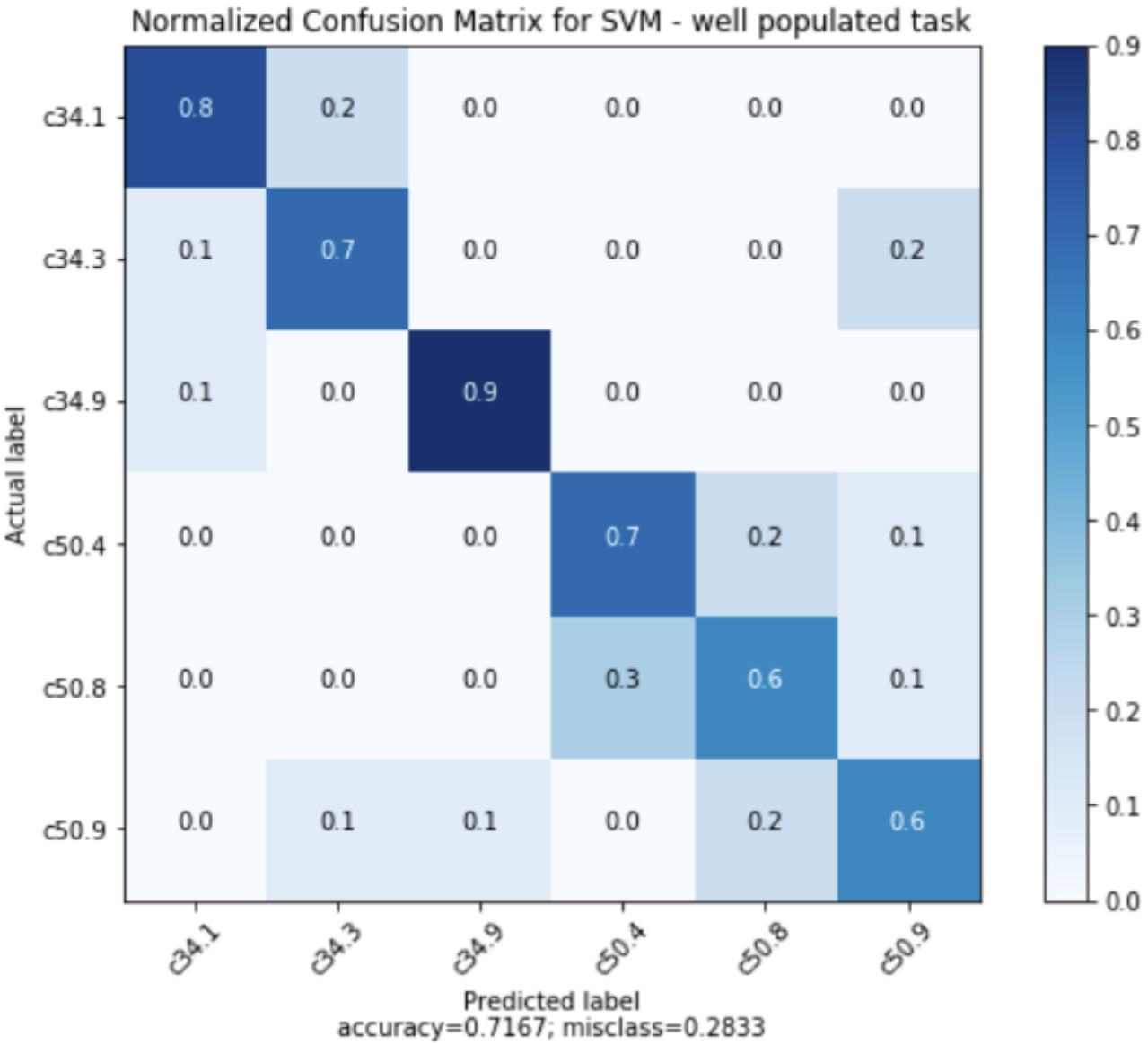

Figure 8

Normalized Confusion Matrix for SVM - well populated task

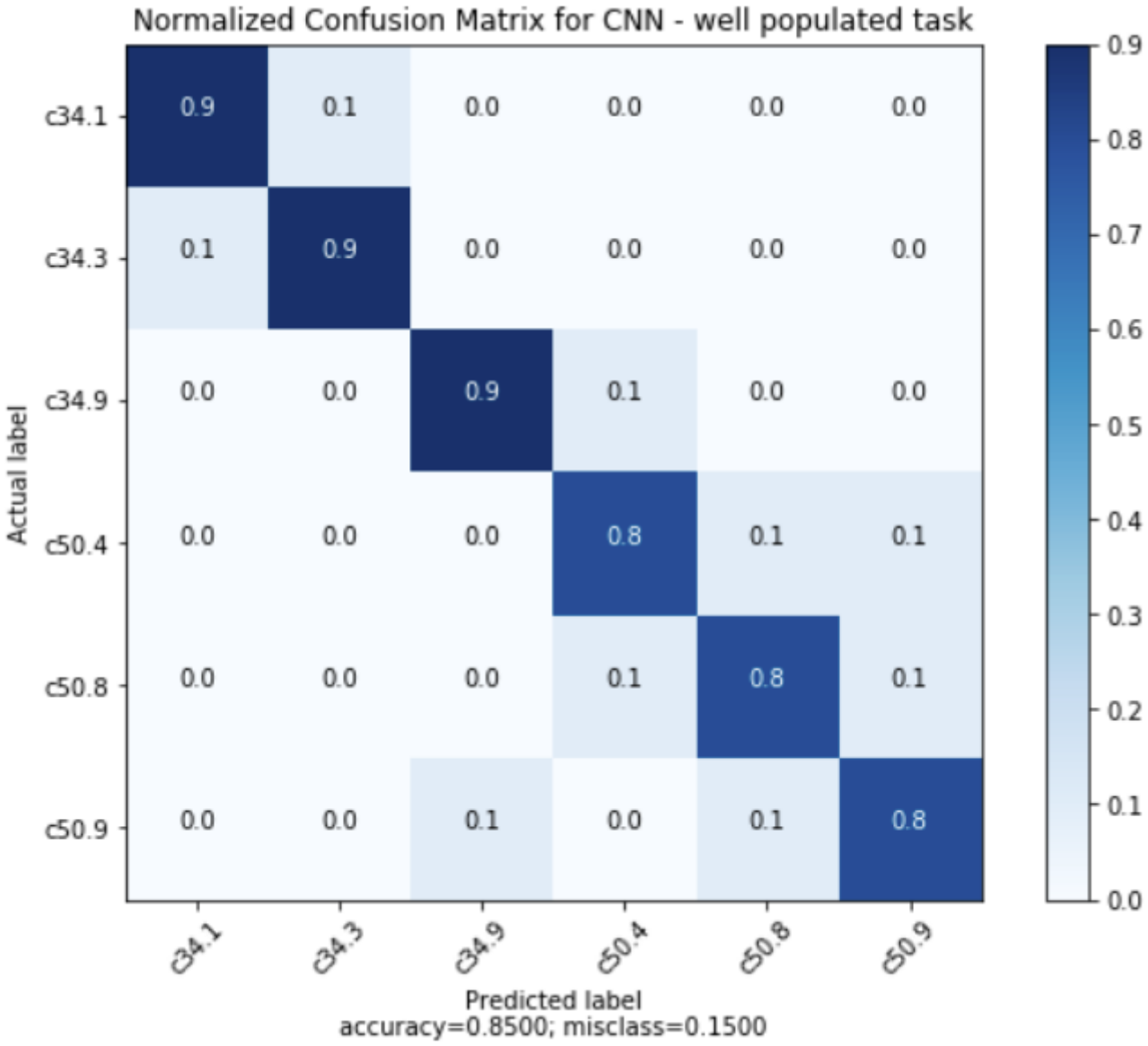

Figure 9

Normalized Confusion Matrix for CNN - well populated task