Volume: 5, Issue: 3
Page: 27-41
2021

International Journal of
Science and Business
Journal homepage: ijsab.com/ijsb

IJSAB International

Medical Text Classification Based on Convolutional Neural Network: A Review

Hazha Saeed Yahia & Adnan Mohsin Abdulazeez

Abstract:
Medical text classification has a significant impact on disease diagnosis, medical research, and the automatic development of disease ontology, acquiring knowledge of clinical results recorded in the medical literature. Hence, medical text classification is challenging because it contains terminologies that describe medical concepts and terminologies. Furthermore, the medical data mostly does not follow natural language grammar; it has inadequate grammatical sentences. The techniques used for text classification give different results comparing to medical text classifications, as extracting text and training sets are different. One of the most significant text classification models in general and medical text classification specifically is CNN-based models. In this paper, many papers on medical text classification have been reviewed, and the details of each article, such as algorithms, or approaches used, databases, classification techniques, and outcomes obtained, are evaluated and outlined thoroughly. Besides, discussions were carried out on all the studied papers, which profoundly influenced medical documents classification.

IJSB
Literature Review
Accepted 30 January 2021
Published 31 January 2021
DOI: 10.5281/zenodo.4483635

Keywords: Text classification, Medical text classification, Neural networks, Convolutional neural networks, CNN architecture.

About Author (s)

Hazha Saeed Yahia (corresponding author), Poly Techniques University, Duhok, Kurdistan Region, Iraq. E-mail: Hazha.yahia@gmail.com

Professor Adnan Mohsin Abdulazeez, Duhok Polytechnic University, Duhok, Kurdistan Region, Iraq.

Introduction

Nowadays, text classification is necessary for business performance, medical diagnosis, and understanding of patients' status, governance, education, and it is essential in almost all sectors. The purpose of text categorization is to decide what categories a text to be categorized belongs to base on each text's attributes (Jiang et al., 2019; Wang et al., 2020). Text classification has numerous applications such as spam detection in email servers, news categorization, medical text analysis, sentiment analysis, hate speech in social media, and many other applications. The text sources are vast, and the source may be from emails, chats, web data, social media, news, user reviews and questionnaires, public speeches, medical and clinical records, and other sources, and all these sources are stored electronically. Moreover, text sources are rich sources of information. However, extracting observation from a text is challenging due to the text's unstructured nature, and the extracting process is time-consuming. It is noticeable that data complexity increases simultaneously with the variability in their structure and format (Aldhoayan and Zhou, 2016; Alam et al., 2020; Zeebaree et al., 2019; Zeebaree et al., 2021; Zeebare et al., 2021). Therefore, text classification is still an unsolved problem (Minaee et al., 2020) in many fields, especially medicine dataset, because of their unstructured nature and high dimensionality. Many techniques and methods are used to analyze text, such as text classification and clustering (Severin et al., 2019). These techniques, when implemented on different sources of text, show notable and good outcomes. While implementing them on medical data and text is still unsatisfied due to the medical dataset's properties (Shabanian et al., 2019).

The medical record registers the medical activity process of a patient's disease development and its medical history and information during the treatment (Al-Doulat et al., 2019). Medical literature is a recorded research document about the diseases and the latest medical results (Abiodun et al., 2018; Zebari et al., 2020; Abdulazeez et al., 2020). Because medical data is heterogeneous and consist of sensitive and unstructured information, the classification of these data requires different methods and techniques(Guo et al., 2020). Moreover, the most widely used model is CNN for medical text classification. CNNs, as a supervised deep learning model, used in various fields such as text and image classifications, disease diagnosis, and other fields (Guohao et al., 2019). The model's ability to learn from complex function representations and end-to-end modes in extensive data and multilevel datasets shows performance, and its prediction basis is appropriate (Wu et al., 2020). This paper aims to review the literature on medical text classification using Convolutional Neural Network (CNN) in various methods. The text classification method, medical text classification, and CNN will be reviewed in the second section. In this field, Section 3 consists of a literature review, contrasting them. Discussing the literature and showing their positive points and drawbacks, followed by sections 4 and 5, finalize the article.

2. Background Theory

In the field of text classification, many kinds of text classification are studied, each of them deals with different types of documents and classes, like topic categorization to unearth mentioned topics, spam detection, and sentiment classification to notice the sentiment generally in available information, especially in social media and news (Sakib et al., 2018; Zeebaree and Abdulazeez, 2020), in this paper medical text classification will be reviewed and studied. As mentioned previously, text classification has two categories; statistical approaches and machine learning approaches (Hindi et al., 2020; Hsu et al., 2020). The Statistical approaches are traditional and mathematical models that work in a similar way to computer programs. To accomplish a valuable classification, the amount of information must be short, and the dimensionality must be reduced (Jiang et al., 2020). Traditional methods cannot classify today's text size and volume. Therefore, the new Machine learning methods for classification should be used (Zhu et al., 2020). Text classification in machine learning is based on learning principles and models, and it is divided into supervised, unsupervised, and semi-supervised text classifications (Abiodun et al., 2018; Sakib et al., 2018), as shown in Figure 1. Supervised learning regulated the classification problem, and the method will be trained and evaluated before starting the actual procedure (Sun and Loparo, 2019; Zhu et al., 2019). Unsupervised learning is used when the data labeled cannot be accessed (Li et al., 2019). Simultaneously, semi-supervised learning is used when the dataset is partially labeled and partially unlabeled (Liu and Chen, 2019).

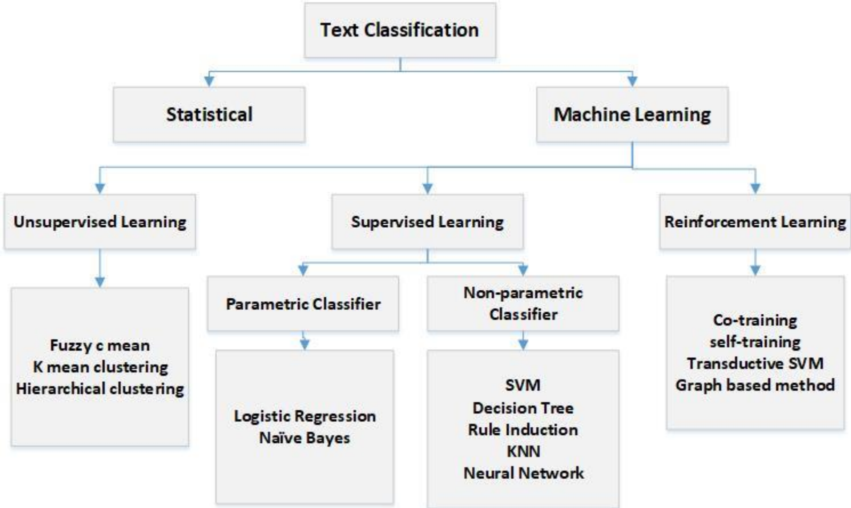

Figure 1. Text classification

a. Convolutional Neural Networks (CNNs)

A convolutional neural network is a deep learning model used in many fields such as natural language processing and text classification, achieved remarkable progress results while using deep learning models (Huang, 2019; Zeebaree and Abdulazeez, 2020)There are other deep learning models such as DNN, and RNN, and each model with the specific structure is the basic model of deep learning, on which many other models are built (Behera et al., 2019; Zeebaree et al., 2019a). Neural networks in general and CNN are considered one of the most potent text classifiers with many properties: robustness, self-learning, and adaptability. Biologically, the neural network represents the human brain, which consists of neurons and synapses organized into layers. The artificial neural networks take advantage of the human neural system and consist of many neurons connected to a centralized system. The neural network consists of three components; Neurons, the neural network' basic unit in which it receives information, perform simple calculations and passes it further to other parts of the neural system (Schmidhuber, 2015; Jahwar & Abdulazeez, 2021; Abdulazeez et al., 2020). The second component in neural networks is Synapses and weights; they connect the neurons and have weight. The last component of neural networks is Bias, which gives more variations of weights to be stored. The neural network architecture consists of an input layer, hiding layers in the middle, and an output layer (Liu et al., 2020). Convolutional Neural Networks (CNNs) are supervised deep learning models used on different problems such as text and image classifications, disease diagnosis, and other areas. It shows success due to the models' ability to learn from complex feature representations and an end-to-end fashion in big data and multileveled datasets (Tensmeyer et al., 2017; Sakib et al., 2018), and its prediction base is acceptable (Zulqarnain et al., 2020). The CNN architecture comprises three main layers; the convolutional layer, the pooling layer, and the fully-connected layers, as shown in Figure 3.

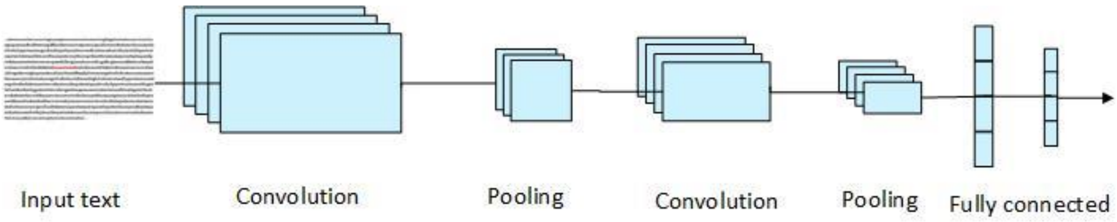

Figure 2 - CNN architecture

The convolutional layer is the most important layer in CNN and its responsibility to learn from the inputs' features (Jacovi et al., 2018). This layer consists of multiple maps; a qualified kernel will have convolved the features maps, then the new convolved feature maps pass to a nonlinear activation function. In this layer, different maps will be created using different kernels (Guo et al., 2017; Behera et al., 2019). In the pooling Layer, the dimension of convolutional layers' features will be reduced, and it considers as the second feature extraction (Guo et al., 2017; Nowak, 2002). Finally, CNN generally has one or more fully-connected layers, which connect all neurons from the previous layer and connects it to the current layer (Guo et al., 2017; Nowak, 2002; Zhang and LeCun, 2015). Recently CNN is widely used in natural language processing (NLP), such as sentiment analysis, text classification, image classification, Intrusion Detection, etc. (Zheng, 2020). Also, it can work on different types of data, such as one-dimension data to multi-dimension data.

b. Medical text classification

Medical data is heterogeneous and have different types and forms. The medical data contains the patient's sensitive data through the medical treatment process, biological data, genetic sequence data, image reports, pathology, cure plans, drug reports, and many other data types(Abdulqader et al., 2020). These data are about the records in presence, but they can be used for the future by analyzing them. Among these data are sensitive data that must be hidden in medical published documents. In short, medical data can be classified into medical images, clinical notes, and some other types of data (Liu et al., 2020). Natural languages contain words, sentences, paragraphs, and each contains elements with explicit and implicit meanings. Understanding these is part of text classification (Zhang and LeCun, 2015). Moreover, this a legal problem because the formation of natural languages is different. The text's situation in the medical text is much complicated than the natural languages (Yu et al., 2019). X-rays, computed tomography (CT), magnetic resonance imaging (MRI), optical coherence tomography (OCT), microscopy, and positron emission tomography are medical image data (PET). Laboratory test results, physician diagnosis, medications, and therapies are included in the clinical text. Physiological measurements (laboratory outcomes, vital signs), demographic details, payment, and insurance information are also included in other medical data (Guo et al., 2017).

c. CNN architecture

Before starting the implementation of any model or algorithm and after data collection, the collected data must be cleaned and classified into training data and test data. These collected data vary in structured, structured, semi-structured, or unstructured, especially with medical data (Song et al., 2019; Sakib et al., 2018). Then the data after cleaning will be fed to the algorithm. This step is done using the training data to teach the machine to increase its predictive accuracy on CNN. When the learned model gets ready, it will be tested using test data, and its predictive accuracy will be checked. The model will be accepted if the predictive accuracy is more than the desired level (Sakib et al., 2018; Abdulazeez et al., 2020). The clean-collected data will be fed to CNN, and the first process is done in the convolution layer. In this layer, the ConvNet, the fundamental unit of the model formed. This layer's parameters focus on using the learnable kernel, which is a tiny spatial dimensionality. Once the information gets into the convolution layer, the layer provides a two-dimensional activation map to the filtered data (Sakib et al., 2018). The convolution layer will check the output of neurons that attacked specific input regions. Then the layer calculated the scalar product between the local weight vector and the input length (Song et al., 2019). Neurons that consist of identical feature maps share the weight, thereby reducing the network's complexness by keeping the number of parameters low. The convolution layer is also responsible for reducing the model's complexity by optimizing its performance by depth, stride, and zero-padding setting (Sakib et al., 2018; Wang and Zhang, 2020; Bargarai et al., 2020). CNN model may have more than one pooling layer which came after the convolutional layer, which deals with the output of convolution layers. The dimensionality of the features will be reduced or cut in the pooling layer. The process of cutting the dimensionality did step by step; therefore, several parameters will be reduced, leading to control overfitting in the model (Shankar et al., 2020; Najim Adeen et al., 2020). Many pooling operations include maximum pooling, average pooling, stochastic pooling, spectral pooling, spatial pyramid pooling, and other operations (Wang and Zhang, 2020). The last layer on CNN is the fully connected layer responsible for taking the transformed vectors. The vectored feature will be integrated to form a model. In this layer, activation functions, such as SoftMax or sigmoid, can make the full connection layer a classifier (Shankar et al., 2020). This layer's outputs are high-level features learned by convolutional layers and then passed to create the output layer.

3. Related Work

Text classification problems have been thoroughly studied and discussed in many research pieces over the last few decades. NLP, is the transformation of a natural language into more functional types. The basic principles in text categorization and document categorization have not changed. This paper reviews the medical text classification literature, starting with clinical data Electronic Health Records, then semantic classification, followed by comments on medical platforms, and finally, the text classification from medical image records. Baker et al. (2016) proposed a CNN model to classify biomedical text from cancer datasets. The dataset is consisting of 1852 biomedical publication abstracts annotated from the hallmarks of cancer. They implemented and evaluated two of each SVM-based methods and CNN variants and trained ten binary classifiers in the multi-label task, one for each hallmark label. The results show that CNN performed better than SVM with an average accuracy of 97.6% and the SVM average accuracy is 94.9%. Authors of (Hughes et al., 2017), proposed a CNN-based approach to category text fragments at the sentence level. Because of the problem of medical phrases, abbreviations, and terminologies, must have a model that solves this problem. Although there are many models and mechanisms, they depend on dictionaries, and dictionaries have limitations, especially with social determinants of health. Then they compare the proposed model with some other Sentence Embeddings, Mean Word Embeddings, and Word Embeddings with BOW (bag-of-word). The dataset set that they use is a collection of 15k clinical research papers representing a wide range of medical subjects. To automatically learn and classify, they used a CNN-based, and they classify the sentences into one of the 26 categories in the dataset. The vectors extracted by using the Word2vec model. The maximum word length is 50. After comparing it with other models, this proposed model's results show that the proposed model has higher accuracy than other models by 0.68% compared with others as (0.28, 0.36, 0.36 and 0.51) repeatedly. In (Nii et al., 2018) they presented a new method known as CNN-based classification methods to classify the medical data in nursing care in Japan to aid and make recommendations. The proposed method works with a word vector representation. Four class labels have been used as (Poor, Fair, Average, and Good), and for extracting the words, Japanese NLP tools have been used and the features extracted by using word2vec. Then the vectors examined the proposed method for finding the most extensive output. The total number of text used is (8313), and they compare the results of the proposed method with other methods like (softmax, kNN, SVM with linear kernels, and SVM with RBF kernels); the results show that in most results, the proposed method is better than the other methods. While (Zeebaree et al., 2018), based on CNN, proposed a new method for selecting features of cancer types in microarray cancer data. By this method, ten cancer datasets were classified and identified. First, the cancer data transformed into an array then the cancer data organized as a matrix-vector. Later, the classification of the vectors applied using CNN. The performance of the proposed method is compared by using mSVM-RFE-iRF and varSeIRF methods. The results of the proposed method are more accurate than the other methods, and its terms of decreasing cancer's genes outperformed the other methods in terms of decreasing cancer's genes.

Yao et al. (2019) presented a new clinical text classification method as clinical text classification Is one of the unsolved problems. Although there are different deep learning methods that show significant feature learning results, the clinical test results are unsatisfied. They proposed a new method for disease classification by combining rule-based feature engineering and knowledge-guided deep learning techniques. Then using CNN for training. They first identify trigger phrases; they examine a patient with obesity with 15 comorbidities. The document labeled as Y for Present, N for unsent, Q for questionable, and U for unmentioned. Then predict classes using trigger phrases using Solt's system. Finally, they use a knowledge-guided CNN for training. The results indicate that CNN models perform better than Perl implementations in the intuitive task, showing that CNNs are superior learning models for more complex tasks. Alternatively, (Yao et al., 2019)proposed another method that combines rule-based features and deep learning techniques for medical text classification as clinical records are essential in EHR data. The electronic medical dataset contains a piece of rich medical information used for medical decisions and research. However, still extracting potential information from these datasets for biomedical researchers is the problem. To stimulate the identification of biomedical entities and minimize information loss, (Gao et al., 2019) proposed HDCNN-CRF method for obtaining a large amount of context information. The dataset has been taken from the NCBI-disease corpus (National Center for Biotechnology Information), 793 Medline abstracts in which 593 used as a training set, and 200 as a testing set. The dataset tested on the proposed method and CRF, and CNN-CRF shows that HDCNN-CRD has higher accuracy than the other two models. The accuracy of HDCNN-CRF is 84.74%, while the accuracy of CRF is 83.80%, and the accuracy of CNN-CRF is 80.26%. While (Yoon et al., 2019) proposed the Text Graph Convolutional (Text GCN) model for finding optimal features of keyword and expressions in medical datasets. Since the word ordering and variation of expression and changing them to natural language processes are challenging medical datasets. Dataset is a Cancer pathology repost dataset taken from SEER registry. The dataset includes 10,000 cancer pathology reports. The accuracy of this model is 95%. Authors

(Xu et al., 2019) proposed a novel genetic mutation prediction tool called CNNBiGRU, which merged CNN with BiGRU, to automate annotated mutation in MSKCC. In the proposed method, CNN used as a feature extraction of semantic information, and BiGRU as the mapper of context semantic structure. The two-channel model trained then fuse the output features of the two to conduct classification. Then they trained and tested their model on a large dataset integrating the two sets of files. They completed the research on CNN, BiGRU, BiGRU, LSTM, Bi-LSTM, LSTM, Bi-LSTM, Cascade neural network based on CNN and BiGRU, BiGRU, and Parallel hybrid neural network based on CNN and BiGRU. The CNN and BiGRU parallel hybrid neural networks display the best performance of the six baseline models. RNN processes, so it functions even better than CNN. The text model works well. CNN is superior to CUDA in its convolution and pooling layers and is more appropriate for text classification. Also, (Qing et al., 2019) presented a novel unified hierarchical neural network method for A standard classification model for the medical text that can be used both in medical records and literature does still not exist. The proposed method is known as AC-BiLSTM, and it is a hybrid neural network combined with CNN. They took the medical text from the datasets of TCM-Traditional Chinese medicine clinical records, CCKS—An open inpatient medical records dataset of China Conference, Hallmarks—a corpus of biomedical publication abstracts annotated hallmarks cancer and AIM—Activating invasion and metastasis. It is a hallmark of cancers. Besides the proposed method, they used CNN, LSTM, RCNN, HAN, SVM, Fasttext, Logistic Regression, and AC-BiLSTM. The results of the proposed method are 89.09%, and 93.75 improved higher than TCM and CCKS. The proposed model indicates the rate of 12.47%, 4.38% compared to CNN on TCM and CCK. Several text classifiers like RCNN, HAN, and AC-BiLSTM, but the findings show that the suggested method counting approach is more effective since they used the attention process in word and sentence levels to retrieve essential terms. Besides, (Liu and Chen, 2019) proposed an algorithm that determines the group to which the disease belongs based on a user's definition of the disorder, which mainly involves informal terminology. Social media is an inexhaustible source of knowledge on a wide variety of subjects. A growing number of people are using social media to discuss and share wellbeing. It is a valuable source that could potentially be used for learning and exploitation. The writing is uncomfortable due to casual comments and harrowing word extraction.

Another source of information in the medical area, is the biomedical figures containing stored information in figures as Radiography scans or stored information in the text as Gene sequence or combination between the two models. Extracting text information from these biomedical figures is essential. Almakky et al. (2019) presented a new approach of text localization in biomedical images. The proposed method is based on a DCNN architecture. They used 500 biomedical figures containing 9308 text regions collected from 288 open-access articles selected from PubMed Central. Banerjee et al. Banerjee et al. (2019) proposed a model for extracting information from free-text reports that include medical images, and they proposed two learning models; CNN Word-Glove and Domain phrase attention-based hierarchical recurrent neural network (DPA-HNN). They used 227,809 radiology reports from the medical center at Duke University, 117,816 reports from Stanford University, and 12,091 reports from Colorado Children's Hospital between 1998-2016. The model trained using 4512 selected reports randomly. The experimental results show that the proposed model trained on a single dataset performed better than the PEFinder, which trained on multi-datasets with an accuracy of 99%. The best F1 score for PE in an adult patient population was 0.99 (DPA-HNN). For pediatrics, the population was 0.99 (HNN), which shows that the deep learning models being trained on adult data demonstrated generalizability to the pediatrics population with comparable accuracy. Because the Electronic Health records and clinical data is one of the most valuable medical data sources, Pan et al. (2020) suggested an innovative approach based on BERT and attention mechanism, called FAMLC-BERT (Feature-level Attention for Multi-label classification on BERT), the suggested method takes unstructured medical history information from Electronic Health Records (EHRs) and predict diseases. ERH data are in text format and are punctuated with misspellings and incorrect grammar. Using their proposed method, they compare many methods, CNN, RNN, and BERT, using Text CNN, fastCNN, Text-RNN, Han, XML-CNN, CRNN, and Bert benchmarks. They found that BERT model has higher performance than other models. Also, the FAMLC-BERT performed fast than other models by 20%. Guo et al. (Guo et al., 2020), one of the medical text classification problems, understands the meaning of some particular words correctly, especially the words with semantic information. Particular words must be captured by using special techniques such as CWU (Context Weight Unit). Although techniques as CWU can extract and highlight the semantic meaning of words, its ability is limited in this area. Therefore, the authors of this article proposed a new mechanism called the self-attention mechanism. The proposed method used the CWU, Region-adjacent embedding (RAE) to classify the Chinese electronic medical records with quantities of popular feature extraction models. The used vocabulary size is (3920), and they text_CNN and BiLSTM with and without RAE. The results show that they are more accurate in most test cases when RAE is using with them.

Pathologists are expected to conduct a large volume of pathology reports yearly. Then information must be extracted from these reports for making the decision or doing researches. Although there are many ways for information extraction, the authors of (Alawad et al., 2018) proposed a training scheme with two stages; Multi-task learning (MTL) and the TM-CNN model. in this first stage, the MTL scheme used to train the MT-CNN model. 942 corpus of 12 classes of cancer has been used with an average length of 469 words, and the results show that the presented method is accurate, with 95%. Also, (Jin et al., 2020) proposed a hybrid deep neural network model and image segmentation model for Chinese text emotion recognition (TBLC-retention) for dataset taking from consumers' comments about cold medicine from an e-commerce platform. The proposed method combines BiLSTM and CNN with attention mechanisms; they obtained the contextual semantic from the comments, then they extract the local semantic features. Several words from the comments are (20,135). The new model performance proposed is higher than the other model, and the accuracy was 99%. In (Ashraf et al., 2020), because the accuracy of classified medical images is very important, training-testing methods are essential, and based on training records, the process of classification and representation will be done. Ashraf et al. (2020) proposed a novel CNN-based algorithm for the image classification by using training-testing methods. Twelve classes have been used; 11 of them are from cancer image archives in a public dataset. Other Twelve classes are taken from a website with open-access, Messidor, for knee images. Each class contains 300 images that are mean the overall images are 3600 images with 12 different classes. 2520 images were used for training, and 1080 used for testing. The proposed algorithm results show that the algorithm's accuracy is 100% with brain and prostate images; it ranges between 97.8 to 99.9 for chest, breast, pancreas, colon, soft tissues, and Esophagus. Then (Ismail et al., 2020) developed CNN-regular pattern discovery model was proposed to detect knowledge from unstructured medical health records related to disease prediction. Medical data from patients with three common diseases were collected, then significant factors were chosen to detect the relationship between specific diseases. Besides, to drive the required understanding, the positive and negative correlated factors analyzed by the co-occurring parameters are selected. 4, 759, 777 medical records are included in the analysis, and the algorithm's precision reached around 80 percent.

Table 1- Overview of the literature in medical text classification using CNN

Ref.		Dataset	Technique(s)	Accuracy	Pros and Cons
(Baker et al., 2016)	2016	1852 biomedical publication	CNN based method	94.9%	Higher accuracy than SVM in 7 tasks out of 10, SVM show better performance in some tasks.
(Hughes et al., 2017)	2017	15k clinical research papers	CNN-based with word2vec	68%	The accuracy is not high, although compared with other methods, it is higher.
(Nii et al., 2018)	2018	8313 text sets from different years (2006, 2007, and 2008)	CNN- based classification method	Not clear	Good classification performance
(Zeebaree et al., 2018)	2018	Cancer datasets	CNN	100% accurate	Very high accuracy and useful od classification of the cancer gene.
(Alawad et al., 2018)	2018	942 corpus of 12 classes of cancer with an average length of 469 words	MTL and TM-CNN	95%	Extracting medical text information is challenging; they could reach a high accuracy by using the proposed models.
(Yao et al., 2019)	2019	22,285 training set and 15,442 Test set	Knowledge-based CNN and Perl implementation	Not clear	Knowledge-based CNN performs better than Perl implementation. However, the knowledge features part does not improve, and instead, it adds noise and unrelated CUIs to previous works.
(Gao et al., 2019)	2019	793 Medline abstracts From NCBI-disease corpus	HDCNN-CRF	84.74%	HDCNN-CRF compared with BiLSTM-CRF is more accurate, and the accuracy of the proposed method is 84.74%, and BiLSTM-CRF is 84.18% is a little less. Nevertheless, the Recall for the second method is higher than the proposed method.
(Yoon et al., 2019)	2019	10,000 cancer pathology reports	Text GCN	95%	The proposed method's accuracy is very high and can be implemented on other document classification and NLP.
(Xu et al., 2019)	2019	Kaggle competition for Redefining Cancer Treatment. 3320 records for the training set and 3320 records for the test set	CNNBiGRU	Accuracy increased by 19.1%	CNNBiGRU compared with the other six models and show the best performance among them. The result would be better if they show the relationship between gene and mutation.
(Qing et al., 2019)	2019	196,537 set of data from TCM, CCKS, Hallmarks, and AIM	BIGRU	89.09% for TCM 93.75% for CCKS 97.73% for AIM 75.72 for Hallmarks	The accuracy in Hallmarks is not fair or less than the other methods because it works fine with medical records than the medical news.

(Liu & Chen, 2019)	2019	19,824 patient description texts from DingXiangyisheng's question and answer module	MSMTC	87.65%	The method proposed would automatically extract words from medical media texts without any usual word references. Trying to deal with the network's terms and non-standard vocabulary.
(Almakky et al., 2019)	2019	DeTEXT dataset	DCNN architecture	91%	Although other methods such as (MSER) shows higher accuracy than the proposed method, they found that there were missed truth boxes in the previous method that the proposed method is covered.
(Banerjee et al., 2019)	2019	More than 367,000 radiology reports	CNN Word-Glove and DPA-HNN	99%	The suggested model improved high accuracy by 99% for single institution datasets, but it has not the same accuracy with multiple datasets
(Pan et al., 2020)	2020	Text from Chinese Electronic Health Records (EHRs) including 300 benchmarks with default pre-trained Word2Vec embedding	FAMLC-BERT	The proposed model is higher than FAMLC-BERT by 10.2%	FAMLC-BERT performed faster than other methods, but they did not show the number of datasets and types of text; besides, they test the proposed model only on Chinses language that the results will be different if compared with other languages such as the English language.
(Guo et al., 2020)	2020	3920	RAE mechanism based on text CNN and BiLSTM	93%	The proposed method has higher accuracy than the other methods. However, the results show that it has not the same accuracy with different sizes. For instance, size 5 and 7 has better accuracy than size 6, without knowing the reason.
(Jin et al., 2020)	2020	Consumer's comment on e-commerce and the size is 20,135	TBLC-retention	99%	Very high accuracy, but the time of training is long.
(Ashraf et al., 2020)	2020	3600 Medical images	CNN-based algorithm	100% for brain and prostate images ranging between 97.8 to 99.9 for chest, breast, pancreas, colon, soft issues, and Esophagus	Very high accuracy with different medical datasets, but the execution time is high

Nugraheni et al. (2020) proposed and experimented with the status of COVID-19 with the CNN model within a short text. The short texts have been collected from news titles after cleansing them using the votes-based mechanisms. The cleaned and labeled new titles are tokenized and vectorized to be transformed into matrix input for the model. then the situation of COVID-19 events classified by using categorial output. 20,431 news titles were collected by searching the keywords “covid” and “corona,” The collected data have been cleaned, and 16,833 data remains. More than half used the remaining data for training (13,116) and the rest for testing (3,717). The results show that the proposed method had an accuracy of about 85%.

4. Discussion

The classification task is one of the most critical issues in machine learning, and many researchers are dealing with text classification in different sources. Many methods, algorithms, and techniques have been used, and one of the most powerful algorithms for dealing with classification in general and text classification in specific is CNN. In Table 1, 17 works on medical text classification based on CNN reviewed in this paper. The accuracy of CNN-based models in 8 works of literature out of 16 is more than 90% (between 90% to 100%), and they have different size of datasets varying from thousands to hundreds of thousands, which ensures that the number of datasets and its type does not affect the accuracy of CNN-based models. Also, CNN-based models show significant results when it deals with the disease datasets, especially cancer datasets. Extracting and classifying cancer datasets have the highest accuracy, from 95% to 100%, and this is higher than other compared methods. The proposed models ensure that the number of datasets does not affect accuracy. It only affects the time of learning and execution. Finally, CNN–based models have less accuracy with the medical texts from websites, clinical research, and patient descriptions because the medical sentiment texts are multidimensional and do not follow standard grammar, making the training process difficult. In contrast, affects the accuracy of the model.

5. Conclusion

This paper has shown an overview of medical text classification based on Convolutional Neural Networks (CNNs). CNNs are supervised deep learning models used on different problems such as text and image classifications, disease diagnosis, and other areas. The literature shows that CNN has different methods, techniques and models to deal with medical text classification and many new algorithms proposed based on CNN. The proposed methods achieved high accuracy results with disease datasets, especially with cancer datasets, and number of datasets not affected the accuracy either they are a large amount of datasets or not. Nevertheless, medical sentiment texts’ results have less accuracy because of their multi-dimensionality. Therefore, classification of the medical sentiment text worthy more additional study and more algorithms and models to be implemented to improve the accuracy of these texts for the importance these texts have in medical literature.

References

Abdulazeez, A. M., Latif, L., & Qader, D. (2020). *Pipelined Parallel Processing Implementation based on Distributed Memory Systems*. *13*(7), 301–312.

Abdulazeez, A. M., Salim, B. W., Zeebaree, D. Q., & Doghramachi, D. (2020). Comparison of VPN Protocols at Network Layer Focusing on Wire Guard Protocol. *International Journal of Interactive Mobile Technologies*, *14*(18), 157–177. https://doi.org/10.3991/ijim.v14i18.16507

Abdulazeez, A. M., Zeebaree, D. Q., & Abdulqader, D. M. (2020). *Wavelet Applications in Medical Images : A Review*. *17265*, 17265–17276.

Abdulqader, D. M., Abdulazeez, A. M., & Zeebaree, D. Q. (2020). Machine learning supervised algorithms of gene selection: A review. *Technology Reports of Kansai University*, *62*(3), 233–244.

Abiodun, O. I., Jantan, A., Omolara, A. E., Dada, K. V., Mohamed, N. A. E., & Arshad, H. (2018). State-of-the-art in artificial neural network applications: A survey. *Heliyon*, *4*(11), e00938.

https://doi.org/10.1016/j.heliyon.2018.e00938

Al-Doulat, A., Obaidat, I., & Lee, M. (2019). Unstructured medical text classification using linguistic analysis: A supervised deep learning approach. *Proceedings of IEEE/ACS International Conference on Computer Systems and Applications, AICCSA*, *2019-Novem*, 1–7. https://doi.org/10.1109/AICCSA47632.2019.9035282

Alam, F., Afzal, M., & Malik, K. M. (2020). Comparative Analysis of Semantic Similarity Techniques for Medical Text. *International Conference on Information Networking*, *2020-Janua*, 106–109. https://doi.org/10.1109/ICOIN48656.2020.9016574

Alawad, M., Yoon, H. J., & Tourassi, G. D. (2018). Coarse-to-fine multi-task training of convolutional neural networks for automated information extraction from cancer pathology reports. *2018 IEEE EMBS International Conference on Biomedical and Health Informatics, BHI 2018*, *2018-Janua*(March), 218–221. https://doi.org/10.1109/BHI.2018.8333408

Aldhoayan, M., & Zhou, L. (2016). An accurate and customizable text classification algorithm: Two applications in healthcare. *2016 IEEE 6th International Conference on Computational Advances in Bio and Medical Sciences, ICCABS 2016*. https://doi.org/10.1109/ICCABS.2016.7802778

Almakky, I., Palade, V., & Ruiz-Garcia, A. (2019). Deep Convolutional Neural Networks for Text Localisation in Figures from Biomedical Literature. *Proceedings of the International Joint Conference on Neural Networks*, *2019-July*(July), 1–5. https://doi.org/10.1109/IJCNN.2019.8852353

Ashraf, R., Habib, M. A., Akram, M., Latif, M. A., Malik, M. S. A., Awais, M., Dar, S. H., Mahmood, T., Yasir, M., & Abbas, Z. (2020). Deep Convolution Neural Network for Big Data Medical Image Classification. *IEEE Access*, *8*, 105659–105670. https://doi.org/10.1109/ACCESS.2020.2998808

Baker, S., Korhonen, A., & Pyysalo, S. (2016). Cancer Hallmark Text Classification Using Convolutional Neural Networks. *BioTxtM 2016*, *BioTxtM*, 1–9. https://cambridgeltl.github.io/cancer-hallmark-cnn/%0Ahttp://www.aclweb.org/anthology/W/W16/W16-51.pdf#page=13%0Apapers3://publication/uuid/C99722D5-A2E6-4B51-A6A7-C4534AB874E9

Banerjee, I., Ling, Y., Chen, M. C., Hasan, S. A., Langlotz, C. P., Moradzadeh, N., Chapman, B., Amrhein, T., Mong, D., Rubin, D. L., Farri, O., & Lungren, M. P. (2019). Comparative effectiveness of convolutional neural network (CNN) and recurrent neural network (RNN) architectures for radiology text report classification. *Artificial Intelligence in Medicine*, *97*(August), 79–88. https://doi.org/10.1016/j.artmed.2018.11.004

Bargarai, F. A. M., Abdulazeez, A. M., Tiryaki, V. M., & Zeebaree, D. Q. (2020). Management of wireless communication systems using artificial intelligence-based software defined radio. *International Journal of Interactive Mobile Technologies*, *14*(13), 107–133. https://doi.org/10.3991/ijim.v14i13.14211

Behera, B., Kumaravelan, G., & Kumar, P. (2019). Performance evaluation of deep learning algorithms in biomedical document classification. *Proceedings of the 11th International Conference on Advanced Computing, ICoAC 2019*, 220–224. https://doi.org/10.1109/ICoAC48765.2019.246843

El Hindi, K., Abu Shawar, B., Aljulaidan, R., & Alsalamn, H. (2020). Improved Distance Functions for Instance-Based Text Classification. *Computational Intelligence and Neuroscience*, *2020*, 1–10. https://doi.org/10.1155/2020/4717984

Gao, M., Wei, H., Chen, F., Qu, W., & Lu, M. (2019). HDCNN-CRF for biomedical text named entity recognition. *Proceedings of the IEEE International Conference on Software Engineering and Service Sciences, ICSESS*, *2019-Octob*(61272369), 191–194. https://doi.org/10.1109/ICSESS47205.2019.9040749

Guo, F., Wu, T., & Jin, X. (2020). An Efficient Method Based on Region-adjacent Embedding for Text Classification of Chinese Electronic Medical Records. *Proceedings - 2020 5th International Conference on Computational Intelligence and Applications, ICCIA 2020*, 183–187. https://doi.org/10.1109/ICCIA49625.2020.00042

Guo, T., Dong, J., Li, H., & Gao, Y. (2017). Simple convolutional neural network on image classification. *2017 IEEE 2nd International Conference on Big Data Analysis, ICBDA 2017*, 721–724. https://doi.org/10.1109/ICBDA.2017.8078730

Guohao, Q., Bin, W., Bai, W., & Baoli, Z. (2019). Competency analysis in human resources using text classification based on deep neural network. *Proceedings - 2019 IEEE 4th International Conference on Data Science in Cyberspace, DSC 2019*, 322–329. https://doi.org/10.1109/DSC.2019.00056

Hsu, C. C., Chang, P. C., & Chang, A. (2020). Multi-Label Classification of ICD Coding Using Deep Learning. *2020 International Symposium on Community-Centric Systems, CcS 2020*. https://doi.org/10.1109/CcS49175.2020.9231498

Huang, T. (2019). A CNN model for SMS spam detection. *Proceedings - 2019 4th International Conference on Mechanical, Control and Computer Engineering, ICMCCE 2019*, 851–861. https://doi.org/10.1109/ICMCCE48743.2019.00195

Hughes, M., Li, I., Kotoulas, S., & Suzumura, T. (2017). Medical Text Classification Using Convolutional Neural Networks. *Studies in Health Technology and Informatics*, *235*(April), 246–250. https://doi.org/10.3233/978-1-61499-753-5-246

Ismail, W. N., Hassan, M. M., Alsalamah, H. A., & Fortino, G. (2020). CNN-based health model for regular health factors analysis in internet-of-medical things environment. *IEEE Access*, *8*, 52541–52549. https://doi.org/10.1109/ACCESS.2020.2980938

Jacovi, A., Shalom, O. S., & Goldberg, Y. (2018). Understanding convolutional neural networks for text classification. *ArXiv*, 56–65. https://doi.org/10.18653/v1/w18-5408

Jahwar, A. F., & Abdulazeez, A. M. (2021). *META-HEURISTIC ALGORITHMS FOR K-MEANS CLUSTERING : A REVIEW*. *17*(7), 1–20.

Jiang, H., Chen, C., Wu, S., & Guo, Y. (2019). Classification of Medical Sensitive Data based on Text Classification. *2019 IEEE International Conference on Consumer Electronics - Taiwan, ICCE-TW 2019*, 2019–2020. https://doi.org/10.1109/ICCE-TW46550.2019.8991726

Jiang, X., Shen, Y., Wang, Y., Jin, X., & Cheng, X. (2020). BaKGraSTeC: A background knowledge graph based method for short text classification. *Proceedings - 11th IEEE International Conference on Knowledge Graph, ICKG 2020*, 360–366. https://doi.org/10.1109/ICBK50248.2020.00058

Jin, Q., Xue, X., Peng, W., Cai, W., Zhang, Y., & Zhang, L. (2020). TBLC-rAttention: A Deep Neural Network Model for Recognizing the Emotional Tendency of Chinese Medical Comment. *IEEE Access*, *8*, 96811–96828. https://doi.org/10.1109/ACCESS.2020.2994252

Li, W., Gao, S., Zhou, H., Huang, Z., Zhang, K., & Li, W. (2019). The automatic text classification method based on bert and feature union. *Proceedings of the International Conference on Parallel and Distributed Systems - ICPADS*, *2019-Decem*, 774–777. https://doi.org/10.1109/ICPADS47876.2019.00114

Liu, K., & Chen, L. (2019). Medical Social Media Text Classification Integrating Consumer Health Terminology. *IEEE Access*, *7*, 78185–78193. https://doi.org/10.1109/ACCESS.2019.2921938

Liu, L., Wu, F. X., Wang, Y. P., & Wang, J. (2020). Multi-receptive-field CNN for semantic segmentation of medical images. *IEEE Journal of Biomedical and Health Informatics*, *24*(11), 3215–3225. https://doi.org/10.1109/JBHI.2020.3016306

Liu, W., Su, J., Mao, Z., Jin, P., Huang, Y., Dou, C., Zhou, L., & Shang, Y. (2020). Research on Text Classification Method of Distribution Network Equipment Fault based on Deep Learning. *Proceedings - 2020 7th IEEE International Conference on Cyber Security and Cloud Computing and 2020 6th IEEE International Conference on Edge Computing and Scalable Cloud, CSCloud-EdgeCom 2020*, 11–16. https://doi.org/10.1109/CSCloud-EdgeCom49738.2020.00012

Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J. (2020). Deep Learning Based Text Classification: A Comprehensive Review. *ArXiv*, *1*(1), 1–43.

Najim Adeen, I. M., Abdulazeez, A. M., & Zeebaree, D. Q. (2020). Systematic review of unsupervised genomic clustering algorithms techniques for high dimensional datasets. *Technology Reports of Kansai University*, *62*(3), 355–374.

Nii, M., Tsuchida, Y., Kato, Y., Uchinuno, A., & Sakashita, R. (2018). Analysis of classification results for the nursing-care text evaluation using convolutional neural networks. *2017 6th International Conference on Informatics, Electronics and Vision and 2017 7th International Symposium in Computational Medical and Health Technology, ICIEV-ISCMHT 2017*, *2018-Janua*, 1–6. https://doi.org/10.1109/ICIEV.2017.8338593

Nowak, R. D. (2002). Nonlinear system identification. In *Circuits, Systems, and Signal Processing* (Vol. 21, Issue 1). https://doi.org/10.1007/BF01211655

Nugraheni, E., Khotimah, P. H., Arisal, A., Rozie, A. F., Riswantini, D., & Purwarianti, A. (2020). *Classifying aggravation status of COVID-19 event from short-text using CNN*. *November*, 240–245. https://doi.org/10.1109/icramet51080.2020.9298674

Pan, Di., Zheng, X., Liu, W., Li, M., Ma, M., Zhou, Y., Yang, L., & Wang, P. (2020). Multi-label Classification for Clinical Text with Feature-level Attention. *Proceedings - 2020 IEEE 6th Intl Conference on Big Data Security on Cloud, BigDataSecurity 2020, 2020 IEEE Intl Conference on High Performance and Smart Computing, HPSC 2020 and 2020 IEEE Intl Conference on Intelligent Data and Security, IDS 2020*, 186–191. https://doi.org/10.1109/BigDataSecurity-HPSC-IDS49724.2020.00042

Qing, L., Linhong, W., & Xuehai, D. (2019). A novel neural network-based method for medical text classification. *Future Internet*, *11*(12). https://doi.org/10.3390/FI11120255

Sakib, S., Ahmed, Jawad, A., Kabir, J., & Ahmed, H. (2018). An Overview of Convolutional Neural Network: Its Architecture and Applications. *ResearchGate*, *November*. https://doi.org/10.20944/preprints201811.0546.v1

Schmidhuber, J. (2015). Deep Learning in neural networks: An overview. *Neural Networks*, *61*, 85–117. https://doi.org/10.1016/j.neunet.2014.09.003

Severin, K., Gokhale, S., & Dagnino, A. (2019). Keyword-based semi-supervised text classification. *Proceedings - International Computer Software and Applications Conference*, *1*, 417–422. https://doi.org/10.1109/COMPSAC.2019.00067

Shabanian, M., Eckstein, E. C., Chen, H., & DeVincenzo, J. P. (2019). Classification of neurodevelopmental age in normal infants using 3D-CNN based on brain MRI. *ArXiv*, *cCMV*, 2373–2378.

Shankar, V., Kumar, V., Devagade, U., Karanth, V., & Rohitaksha, K. (2020). Heart Disease Prediction Using CNN Algorithm. *SN Computer Science*, *1*(3), 1–8. https://doi.org/10.1007/s42979-020-0097-6

Song, P., Geng, C., & Li, Z. (2019). Research on Text Classification Based on Convolutional Neural Network. *Proceedings - 2nd International Conference on Computer Network, Electronic and Automation, ICCNEA 2019*, 229–232. https://doi.org/10.1109/ICCNEA.2019.00052

Sun, Y., & Loparo, K. (2019). Information extraction from free text in clinical trials with knowledge-based distant supervision. *Proceedings - International Computer Software and Applications Conference*, *1*, 954–955. https://doi.org/10.1109/COMPSAC.2019.00158

Tensmeyer, C., Saunders, D., & Martinez, T. (2017). Convolutional Neural Networks for Font Classification. *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, *1*, 985–990. https://doi.org/10.1109/ICDAR.2017.164

Wang, L., & Zhang, B. (2020). Fault Text Classification Based on Convolutional Neural Network. *2020 IEEE 7th International Conference on Industrial Engineering and Applications, ICIEA 2020*, 937–941. https://doi.org/10.1109/ICIEA49774.2020.9101960

Wang, S., Pang, M., Pan, C., Yuan, J., Xu, B., Du, M., & Zhang, H. (2020). Information Extraction for Intestinal Cancer Electronic Medical Records. *IEEE Access*, *8*, 125923–125934. https://doi.org/10.1109/ACCESS.2020.3005684

Wu, F., Lin, Y.-C., & Wu, Y.-H. (2020). *CNN-Based Medical Device Design for the Smart Pillbox With Face Recognition*. 607–611. https://doi.org/10.1109/iccst50977.2020.00125

Xu, J., Zheng, X., & Jiang, M. (2019). Gene Mutation Classification Using CNN and BiGRU Network. *9th International Conference on Information Science and Technology, ICIST 2019*, 397–401. https://doi.org/10.1109/ICIST.2019.8836846

Yao, L., Mao, C., & Luo, Y. (2019). Clinical text classification with rule-based features and knowledge-guided convolutional neural networks. *BMC Medical Informatics and Decision Making*, *19*(Suppl 3). https://doi.org/10.1186/s12911-019-0781-4

Yoon, H. J., Gounley, J., Young, M. T., & Tourassi, G. (2019). Information Extraction from Cancer Pathology Reports with Graph Convolution Networks for Natural Language Texts. *Proceedings - 2019 IEEE International Conference on Big Data, Big Data 2019*, 4561–4564. https://doi.org/10.1109/BigData47090.2019.9006270

Yu, Y., Li, M., Liu, L., Li, Y., & Wang, J. (2019). Clinical big data and deep learning: Applications, challenges, and future outlooks. *Big Data Mining and Analytics*, *2*(4), 288–305. https://doi.org/10.26599/BDMA.2019.9020007

Zebari, D. A., Zeebaree, D. Q., Abdulazeez, A. M., Haron, H., & Hamed, H. N. A. (2020). Improved Threshold Based and Trainable Fully Automated Segmentation for Breast Cancer Boundary and Pectoral Muscle in Mammogram Images. *IEEE Access*, *8*, 203097–203116. https://doi.org/10.1109/access.2020.3036072

Zeebaree, D., & Abdulazeez, A. M. (2020). *Evaluating Data Mining Classification Methods Performance in Internet of Things Applications*. *December*, 10–25. https://doi.org/10.30880/jscdm.2020.01.02.002

Zeebaree, D., Mohsin Abdulazeez, A., Asaad Zebari, D., Haron, H., & Nuzly Abdull Hamed, H. (2021). Multi-Level Fusion in Ultrasound for Cancer Detection based on Uniform LBP Features. *Computers, Materials & Continua*, *66*(3), 3363–3382. https://doi.org/10.32604/cmc.2021.013314

Zeebaree, D. Q., Haron, H., & Abdulazeez, A. M. (2018). Gene Selection and Classification of Microarray Data Using Convolutional Neural Network. *ICOASE 2018 - International Conference on Advanced Science and Engineering*, *February*, 145–150. https://doi.org/10.1109/ICOASE.2018.8548836

Zeebaree, D. Q., Haron, H., Abdulazeez, A. M., & Zebari, D. A. (2019a). Machine learning and Region Growing for Breast Cancer Segmentation. *2019 International Conference on Advanced Science and Engineering, ICOASE 2019*, *April*, 88–93. https://doi.org/10.1109/ICOASE.2019.8723832

Zeebaree, D. Q., Haron, H., Abdulazeez, A. M., & Zebari, D. A. (2019b). Trainable Model Based on New Uniform LBP Feature to Identify the Risk of the Breast Cancer. *2019 International Conference on Advanced Science and Engineering, ICOASE 2019*, *April*, 106–111. https://doi.org/10.1109/ICOASE.2019.8723827

Zhang, X., & LeCun, Y. (2015). *Text Understanding from Scratch*. http://arxiv.org/abs/1502.01710

Zheng, W. F. (2020). Intrusion Detection Based on Convolutional Neural Network. *Proceedings - 2020 International Conference on Computer Engineering and Application, ICCEA 2020*, 273–277. https://doi.org/10.1109/ICCEA50009.2020.00066

Zhu, Yinglin, Zheng, W., & Tang, H. (2020). Interactive Dual Attention Network for Text Sentiment Classification. *Computational Intelligence and Neuroscience*, *2020*. https://doi.org/10.1155/2020/8858717

Zhu, Yishun, Jin, X., & Li, L. (2019). Automatic conversion of electronic medical record text for openEHR based on semantic analysis. *Proceedings - 10th International Conference on Information Technology in Medicine and Education, ITME 2019*, 35–39. https://doi.org/10.1109/ITME.2019.00019

Zulqarnain, M., Ghazali, R., Hassim, Y. M. M., & Rehan, M. (2020). A comparative review on deep learning models for text classification. *Indonesian Journal of Electrical Engineering and Computer Science*, *19*(1), 325–335. https://doi.org/10.11591/ijeecs.v19.i1.pp325-335

Cite this article:

Hazha Saeed Yahia & Adnan Mohsin Abdulazeez (2021). Medical Text Classification Based on Convolutional Neural Network: A Review. *International Journal of Science and Business*, *5*(3), 27-41. doi: https://doi.org/10.5281/zenodo.4483635

Retrieved from http://ijsab.com/wp-content/uploads/683.pdf

Published by

IJSAB International