electronics

MDPI

Article

Deep Learning Based Biomedical Literature Classification Using Criteria of Scientific Rigor

Muhammad Afzal [1,*], Beom Joo Park [2], Maqbool Hussain [1] and Sungyoung Lee [2,*]

[1] Department of Software, Sejong University, Seoul 05006, Korea; maqbool.hussain@sejong.ac.kr
[2] Ubiquitous Computing Lab, Department of Computer Science and Engineering, Kyung Hee University, Seocheon-dong, Giheung-gu, Yongin-si, Gyeonggi-do 446-701, Korea; pbj@oslab.khu.ac.kr
* Correspondence: mafzal@sejong.ac.kr (M.A.); sylee@oslab.khu.ac.kr (S.L.); Tel.: +82-31-201-2514 (S.L.)

Received: 1 July 2020; Accepted: 3 August 2020; Published: 5 August 2020

Abstract: A major blockade to support the evidence-based clinical decision-making is accurately and efficiently recognizing appropriate and scientifically rigorous studies in the biomedical literature. We trained a multi-layer perceptron (MLP) model on a dataset with two textual features, title and abstract. The dataset consisting of 7958 PubMed citations classified in two classes: scientific rigor and non-rigor, is used to train the proposed model. We compare our model with other promising machine learning models such as Support Vector Machine (SVM), Decision Tree, Random Forest, and Gradient Boosted Tree (GBT) approaches. Based on the higher cumulative score, deep learning was chosen and was tested on test datasets obtained by running a set of domain-specific queries. On the training dataset, the proposed deep learning model obtained significantly higher accuracy and AUC of 97.3% and 0.993, respectively, than the competitors, but was slightly lower in the recall of 95.1% as compared to GBT. The trained model sustained the performance of testing datasets. Unlike previous approaches, the proposed model does not require a human expert to create fresh annotated data; instead, we used studies cited in Cochrane reviews as a surrogate for quality studies in a clinical topic. We learn that deep learning methods are beneficial to use for biomedical literature classification. Not only do such methods minimize the workload in feature engineering, but they also show better performance on large and noisy data.

Keywords: healthcare; deep learning; evidence-based medicine; biomedical literature; health information management

1. Introduction

In the practice of modern-age medicine, providing appropriate evidence for clinical decisions plays a vital role in increasing the reliability of the system while practicing evidence-based medicine [1]. However, the classification of a vast set of medical documents is daunting and time consuming. In order to solve these problems, recent research on the classification of documents through machine learning is being actively carried out. Additionally, machine learning algorithms have presented a special use in data mining applications, especially where it is difficult for the human to understand and model the domain [2]. Artificial neural networks appear to advance further the area of biomedical literature mining and classification [3,4]. One of the advantages of deep learning over shallow machine learning is the automatic feature engineering. However, the main issue for deep learning models is the acquisition of pre-annotated data required for training and testing its performance.

This study aims at building a deep learning binary classifier for the identification of quality studies in the biomedical literature to prove its superiority to the promising machine learning approaches historically used for the same problem. To achieve this objective, we design deep learning based on a multi-layer feed-forward artificial neural network also called multi-layer perceptron (MLP) model

tied with automatic engineering of features created from the text in title and abstract of biomedical documents related to Kidney disease. To obtain studies that are scientifically rigorous, we run a set of queries related to Kidney disease on the Cochrane database. The same queries are executed using PubMed to get labels for studies that not scientifically rigorous. The proposed model secures a better score as compared to shallow machine learning algorithms.

The main contributions of this work are made in the following areas: (i) the identification and preparation of data using Cochrane and PubMed databases, (ii) the design of an artificial neural network (ANN)-based deep learning model, and (iii) the empirical as well as qualitative insights of deep learning in comparison to shallow machine learning algorithms.

2. Background and Related Work

Over more than two decades, evidence-based medicine has rightfully become part of the fabric of modern clinical practice and has contributed to many advances in healthcare [5]. To work in the domain of evidence-based medicine, clinicians need to keep up with current research findings. However, the issue is that they face challenges in accessing scientifically sound studies in an accurate and timely manner. Currently, the biomedical knowledge has increased many-fold, which spurred interest in the techniques like text mining and natural language processing (NLP) to utilize when dealing with the vast body of biomedical articles [6]. In addition, the biomedical literature is growing by more than one million studies per year, which makes it difficult for researchers to keep pace with the new knowledge published on different topics [3]. These rapid advancements in the domain make it increasingly time consuming for the researchers to critically appraise research studies to find the evidence of clinical impact [7]. The issue is lacking automated and reliable computerized methods to support researchers in recognizing studies of scientific rigor. With the advent of data-driven approaches, we have the opportunity to apply machine- and deep learning algorithms to get autonomous access and appraisal of the biomedical literature.

2.1. Use of Machine Learning and Deep Learning for Biomedical Literature Classification

As a subfield of artificial intelligence, machine learning allows machines to learn from data by designing and developing intelligent algorithms and techniques. That is, the machine learns the pattern and characteristics of data, evaluates and predicts new data based on it, and enables us to utilize it. Machine learning can be divided into supervised learning and unsupervised learning. In a supervised learning model, the algorithm learns from the categorized data set and provides the criteria that the algorithm can use to evaluate accuracy in learning data. Conversely, the unsupervised learning model provides unclassified data, and the algorithm tries to understand the data by extracting its characteristics and patterns. Among the classification algorithms, there is a range of classifiers. For instance, the Naïve Bayes classifier is a typical generative classifier and is regarded as a special case of Bayesian Network classifiers [8]. The support vector machine (SVM) algorithm learns how important each training data point is to distinguish the decision boundaries between the two classes during learning. In general, some of the training data, only the data points located at the boundary between the two classes, influence the decision boundary. These data points are called support vectors.

Deep Learning models are a special kind of machine learning that allows computational models to learn data with multiple levels of abstraction through multiple processing layers. Deep learning discovers complex structures in big datasets by using the backpropagation algorithms to show how a machine should alter its inner parameters using the representation in the past layer to calculate the representation in each layer. There are three important types of neural networks used in deep learning models: convolutional neural networks (CNN), recurrent neural networks (RNN), and multi-layer artificial neural networks (ANN). Deep convolutional networks have made breakthroughs in image, video, voice, and audio processing, while recurrent networks have shed light on sequential information such as text and voice [8]. In contrast, multi-layer ANN is a suitable option for the classification of textual data structured in a tabular form.

NLP, data mining, and machine and deep learning techniques work together to classify and discover patterns in the text of biomedical documents automatically. The primary objective of text mining is to allow users to obtain data from textual resources and deal with such activities as retrieval, classification, and summarization [9]. Anderlucci, Laura, et al. provide a detailed comparison of the shallow, ensemble, and deep learning methods used for classification of textual data [10]. Initially invented for computer vision, CNN models have subsequently been shown to be useful for NLP and have achieved excellent results in semantic parsing [11]. A CNN-based deep learning model [12] by Del Fiol et al. was trained using a large, noisy dataset of PubMed citations with title and abstract as features and obtained comparatively better results as compared to competitors that include PubMed's Clinical Query Broad treatment filter and McMaster's text word search strategy. Using supervised machine learning methods, Sarkar et al. develop a model for identifying quality articles using data features of title, abstract, and others [13]. Bian et al. proposed a machine learning-based high-impact classifier [14] trained on a set of different features and claimed to outperformed the high-quality Naïve Bayes classifier proposed by outperforms Kilicoglu et al. 's [7]. Afzal et al. built compared different machine learning algorithms and learned to choose a support vector machine (SVM) based model due to its higher performance [15]. As an extension to this work, we propose an MLP-based binary classifier and compare its performance with shallow and ensemble machine learning methods.

2.2. Use of Cochrane Reviews for Annotation of Scientifically Rigor Studies

Cochrane Collaboration is an international organization that prepares, maintains, and offers available systematic reviews of health care interventions' benefits/risks. The Cochrane Library is commonly considered to be the best source of credible healthcare evidence [16]. Systematic reviews use a transparent and systematic process to construct study questions, look for studies, evaluate their quality, and synthesize qualitative or quantitative results [17]. The Cochrane Database of Systematic Reviews is the world's most vibrant resource of meta-analysis, with 54 active organizations responsible for organizing, advising, and publishing systematic reviews. In this paper, we utilize studies cited in Cochrane reviews as a surrogate for quality studies. Because of the recognition of Cochrane reviews at a global scale as the best standard in evidence-based medicine, we, therefore, adopted it for the classification of primary documents referenced in the reviews as high-impact evidentiary articles.

3. Methods

We propose an MLP model for evaluating studies of scientific rigor. As depicted in Figure 1, our proposed method consists of two steps. Step 1 is dedicated to acquiring identifiers of studies through queries from the sources and apply filters to remove duplications in the data. Step 2 processes the deduplicated data to get the text of titles and abstracts, which are then preprocessed to create feature vectors to apply the proposed MLP model to classify the studies into scientific rigor and non-rigor.

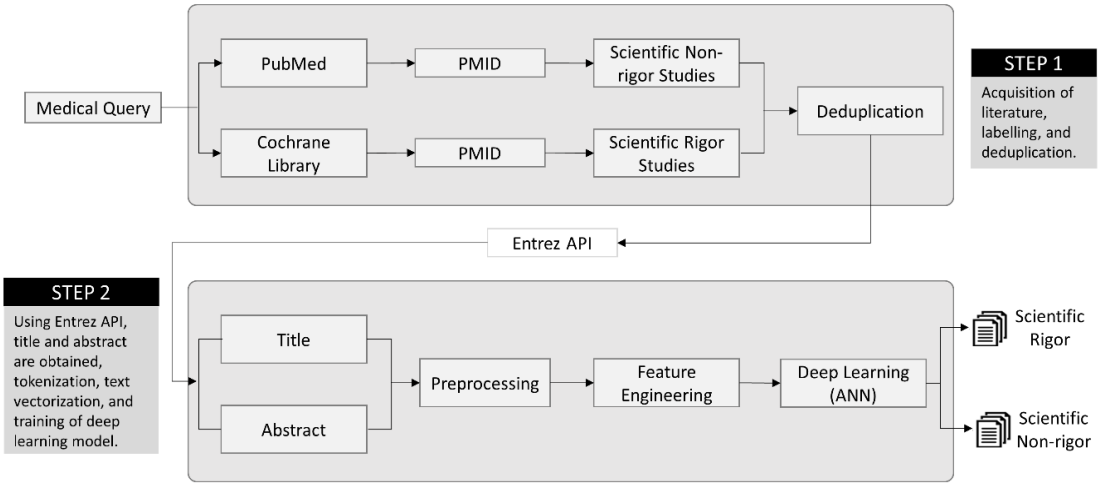

Figure 1. The high-level architecture of the proposed method for finding high-impact studies in biomedical literature.

3.1. Step 1—Preparation of Datasets

To collect high-quality studies, we use the Cochrane Library by executing a general-purpose query of Kidney disease and obtain the identifiers called PMID (PubMed Identifiers) for all the retrieved articles. The same query is executed using PubMed and the PMIDs for all the retrieved articles are obtained. The records obtained from the Cochrane are labeled with "scientific rigor" class, and the records obtained from PubMed are labeled with "scientific non-rigor." In the step of deduplication, all duplicated items are removed from the collected PMIDs in the PubMed dataset in order to avoid the chance of assigning two labels to the same study. Identifiers in both the datasets are uploaded using NCBI Entrez Programming Utilities (eUtils) [18]. The eUtils is a service of NCBI that provides access to a total of 50 databases via a web interface, a public FTP (file transfer protocol) site [19]. Using search and fetch functions of the API, we retrieve two data features (title and abstract) from each article. At this stage, the title and abstract are in the original format consisting of texts as written in the publications. The combined dataset holds 7958 records, out of which 1083 are classed as scientific rigor and the rest as scientific non-rigor.

3.2. Step 2—Development of Deep Learning Model

To achieve the optimal design for MLP, we employ the Auto Model extension of RapidMiner, which provides a graphical visual environment for the convenience of designing a faster and better model of automated classification and prediction [20]. To learn the classification model, the dataset is passed through several internal steps that include preprocessing, feature extraction, and feature engineering, as shown in the detailed diagram (Figure 2) of step 2 of our proposed methodology.

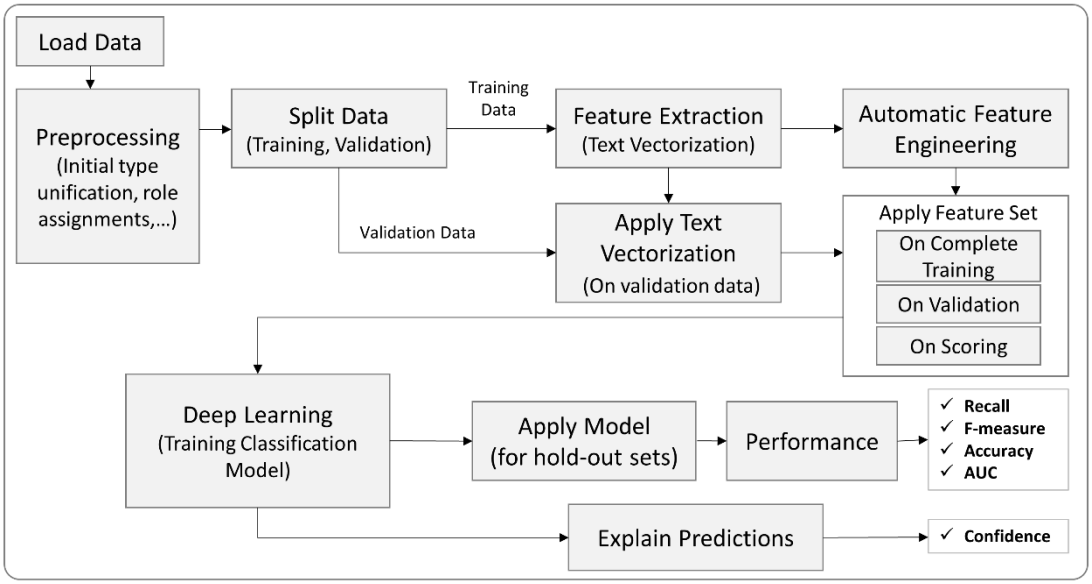

Figure 2. Abstract description of the proposed deep learning process model.

3.2.1. Preprocessing

Preprocessing consists of multiple sub steps such as the role setting and the transformation of the initial types. In the role setting step, all attributes' roles are changed to regular except the class attribute, which is set to the 'label' role. In the transformation of the initial types step, all the text columns are transformed into polynomial columns. After initial preprocessing, the data is checked for missing values and filled where applicable; for instance, the missing numeric values are replaced with the average value. Finally, the values are filtered based on the no_missing_labels parameter that keeps only those records that do not have a missing value in the special attribute with the label role. The data is then split into training set and validation set (holdout data) with a ratio of 6:4, respectively.

3.2.2. Feature Extraction

Technically, this step performs tokenization, changing the case to lower, and calculating TF-IDF (term frequency–inverse document frequency) values for each token across all the records. The input text in the nominal form is transformed into a vectorized format using TF-IDF. The TF-IDF is a statistical value which reflects how important a word is to a document in a corpus. This step helps later in the feature selection based on the importance determined by TF-IDF value.

3.2.3. Automatic Feature Engineering

The automatic feature engineering is a robust utility of RapidMiner that uses a deep learning MLP model internally for the selection of a subset of features from a full set of features. The MLP comes with default parameters that are optimized for finding the best feature sets. Taking all records as a training dataset, the MLP is trained to obtain the final features on the default parameter settings provided by the Auto Model extension of RapidMiner.

3.2.4. Classification Model

After obtaining the best feature sets through automatic feature engineering, a classification model was used to identify the scientific rigor studies. The classification model is the same deep learning MLP model as used for feature engineering. The parameter settings, as described in Table 1, are the optimized default values provided by the RapidMiner.

Table 1. Parameter settings of the proposed deep learning (multi-layer perception (MLP)) classification model.

Parameter	Value	Description
activation	Rectifier	Rectifier Linear Unit is the activation function used by the hidden layers to choose the maximum value of the input.
hidden layer sizes	Two layers each with a size of 50 neurons	The number of hidden layers is limited to only two layers for the reason of the simplicity and efficiency of the model.
local random seed	1992	Local random numbers are the pseudo-random number assigned initially to start the network.
epochs	10.0	The number of times the dataset should be iterated.
epsilon	1.0×10^8	Similar to the learning rate, it allows forward progress.
rho	0.99	rho is the "Gradient moving average decay factor" used for the learning rate decay over each update.
L1	1.0×10^5	It is a regularization method that constrains the absolute value of the weights.
L2	0.0	It is a regularization method that constrains the sum of the squared weights.
missing values handling	Mean Imputation	Missing values are replaced with the mean value.

3.2.5. Performance and Explanation

The trained classification model is applied on hold-out datasets to get the performance in the form of widely used matrices in the AI domain that include accuracy, recall, precision, F-measurement, and AUC. In addition to these metrics, we also explain the model's predictions in the form of confidence value. The higher confidence value of a prediction shows model trust on a classification of a study leads the users to accept the output of the model with a firmer belief.

3.3. Model Selection

To prove the hypothesis that deep learning model can perform better than the shallow machine learning algorithms, we choose four well-known algorithms which have been experimented with in multiple studies [7,13,15] that include Naïve Bayes (NB), Decision Tree (DT), Support Vector Machine (SVM), and Gradient Boosted Trees (GBT). A brief description of parameter settings for each algorithm is provided in Table 2.

Table 2. Descriptions of parameter settings of shallow machine algorithms used for comparison with deep learning.

Algorithm	Parameter Settings
Naïve Bayes	laplace correction (True)
Decision Tree	criterion (gain_ration), maximum depth (10), apply pruning (True)
Support Vector Machine	svm-type (c-SVC), kernel type (rbp), c (0.0)
Gradient Boosted Trees	number of trees (100), maximum depth (10), learning rate (0.01), sample rate (1.0)

We compare the performance of these algorithms with proposed MLP on the basis of an accumulated score obtained from summing up the widely used performance metrics such as accuracy, recall, F1-measure, and AUC (area under the ROC curve). In order to evaluate the variations, we check

to compare the performance of each algorithm individually on title and abstract and then finally on the combination of them, as shown in three compartments of Table 3.

Table 3. Results for learning methods on hold-out datasets.

Data Set: Title					
Method	ACC (%)	AUC	Recall (%)	F1 Score (%)	overall score
NB	75.9	0.696	86.7	49.5	212.796
MLP	94.1	0.942	81.2	78.9	255.142
DT	91	0.679	36.2	52.1	179.979
SVM	94.8	0.968	71.8	79	246.568
GBT	92.74	0.941	70.87	72.62	237.171
Data Set: Abstract					
Method	ACC (%)	AUC	Recall (%)	F1 Score (%)	overall score
NB	90.1	0.882	89.2	70.9	251.082
MLP	92.1	0.967	89.1	75	257.167
DT	92.7	0.733	56.5	67.6	217.533
SVM	95.7	0.974	79.0	83.2	258.874
GBT	94.74	0.969	76.84	79.67	252.219
Data Set: Title + Abstract					
Method	ACC (%)	AUC	Recall (%)	F1 Score (%)	overall score
NB	88.2	0.830	87.4	66.9	243.33
MLP	97.3	0.993	95.1	90.4	283.793
DT	91.9	0.738	51.6	63.4	207.638
SVM	95.9	0.985	74.2	83.2	254.285
GBT	95.16	0.974	83.83	82.5	262.464

We can see the performance of each model in contrast individually on the accuracy, AUC, recall, and F1-measure scores as well as the overall score. Applying the proposed MLP using the dataset with the title and abstract has the highest overall score. Comparing models on to title feature alone, we can observe the highest overall score (value: 255) is slightly less than the highest overall score (value: 257) of models trained on the abstract feature. When the two features are combined, the overall score for almost all algorithms has been increased. We can observe the highest overall score (value: 283) of MLP trained on the combination of title and abstract is considerably better in each of the machine learning algorithms. The second-highest performer, i.e., GBT overall score (value: 262), is about 10% less than the MLP score. Not only the overall score, but also the MLP performs better at each metric, i.e., the accuracy (97.3%), AUC (0.993), recall (95.1%), and F1-measure (90.4%) are higher by about 2, 1, 7, and 7%, respectively, than the second-best performer. Based on these measurements, we conclude that the proposed deep learning model is a better performer on the given dataset; we choose the MLP to test on datasets from a selected clinical domain.

4. Results and Discussion

We apply the MLP model to evaluate the performance of the unseen test data retrieved with a set of real-world queries executed for two domains: kidney disease (same domain used in training) and cancer disease (different domain). As shown in Figure 3, the process steps used for data manipulations are similar to the steps used at the training stage of the model, i.e., the preprocessing, feature extraction, and automatic engineering. The already trained MLP classification model is loaded to apply to the feature sets created from the data retrieved with queries.

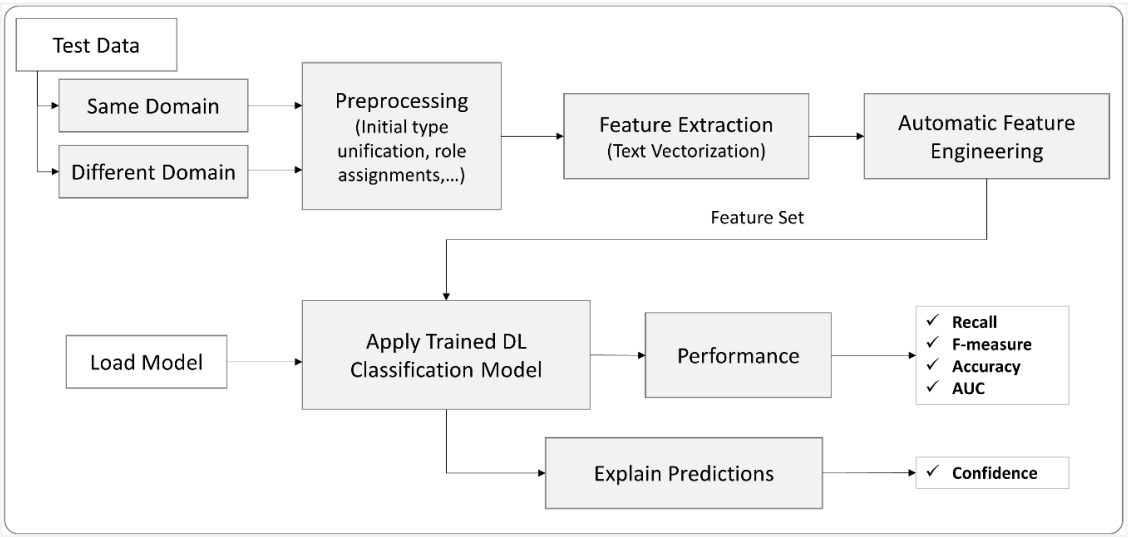

Figure 3. The process steps of application of the trained deep learning model to the test dataset.

4.1. Scenario 1: Results of Same Domain Test Queries

Hypothesis 1. *The proposed deep learning model yields equivalent or marginally lower accuracy for unseen test dataset retrieved with real world queries from the same domain. We define four queries and collect the articles by running each query on the PubMed database. The defined queries and the number of articles retrieved against each query are shown in Table 4.*

Table 4. Same domain test queries and number of articles.

	Query	Number of Articles	
		High-Quality	General
Q1	Chronic Kidney Disease	1051	5527
Q2	Diabetic Kidney	1045	5784
Q3	Kidney Transplantation	1047	6732
Q4	Acute Kidney	902	5398

The results for each query are shown in Table 5. We can observe that there is the least variation in the performance of the model on each query. Additionally, each metric performance is marginally lower than the performance on a hold-out dataset, which proves our Hypothesis 1.

Table 5. The result of evaluating performance of the proposed model on test dataset collected through queries from the same domain i.e., Kidney.

Query	ACC (%)	AUC	Recall (%)	F1 Score (%)
Q1	89.78	0.911	95.82	94.03
Q2	91.17	0.877	96.53	94.87
Q3	90.98	0.875	95.70	94.79
Q4	88.77	0.852	95.97	93.61

4.2. Scenario 2: Different Domain Test Queries Results

Hypothesis 2. *The proposed deep learning model yields equivalent or marginally lower (not less than 10%) accuracy for unseen test dataset retrieved with real world queries from a different domain.*

To test this hypothesis, we similarly collected data as we collected for other queries, however, for a different domain (cancer disease). We collected 1022 high-quality literature data and 6569 general literature data. This cross-domain results in evaluating the proposed model are shown in Table 6.

Table 6. The result of evaluating performance of the proposed model on test dataset collected through queries for a different domain (i.e., cancer).

Domain	ACC (%)	AUC	Recall (%)	F1 Score (%)	Number of Datasets	
					High-Quality	General
Cancer	88.08	0.781	99.12	93.5	1022	6569

We can observe that the performance is slightly lower than the average score of the same domain, however it is nearly equivalent to query 4, which has the lowest score in the four queries of the same domain. The marginally lower performance (not less than 10%) than the score on the hold-out dataset, proves Hypothesis 2.

4.3. Significant Findings

To our knowledge, this is the first study to use Cochrane Systematic Reviews for training deep learning techniques to identify scientifically sound studies in the biomedical literature in the Kidney domain. Besides, our proposed deep learning model performed reasonably well compared with state-of-the-art machine learning approaches. Beyond higher accuracy, we noticed that the proposed deep learning model performed well for unknown test results obtained with the real-world queries. It has shown excellent performance when experiments are performed through multiple queries within the same domain. In addition, our model performed considerably well in other domains, like cancer. Besides, the results were consistent for testing datasets across different queries. The minimum variation across different queries and different domains, qualify our proposed model to be useful to use for identification of high-quality medical articles in the biomedical literature.

4.4. Comparison with Prior Work

There have been numerous attempts to recognize high-quality medical literature that has been automated using computing techniques. A comparison of previous studies and our proposed model is shown in Table 7. Although the comparison is thematic as the overall objective of the mentioned studies is the same, i.e., the identification of high-impact studies in the biomedical literature, it provides a perspective on the superiority of the deep learning model over shallow machine learning methods. From a method perspective, the study conducted by Del Fiol et al. [12] is closer to our study as they have also used a deep learning method in their experiments. The difference was that they used deep learning based on CNN, while the proposed method used MLP—a deep learning based on multi-layer feed-forward neural networks. They utilized clinical queries data as a surrogate for high-impact studies while we used Cochrane citations as a surrogate for scientific rigor (high-impact) studies. Both studies used titles and abstracts as features for the experiment. In both cases, the data were noisy because of the existence of false positives in the dataset. The recall in both cases is similar to about a 1% difference. However, there is a huge difference in F-measure due to the higher precision of our approach. One of the possible reasons is the noise in data, i.e., the number of false positives. Another possibility is that using Cochrane citations as a surrogate for high-impact studies may be more impactful as compared to clinical queries or clinical hedges.

Table 7. Performance comparison of the proposed model and past research.

Research Work	AI Method Used	Performance Metric	Datasets and Features
Del Fiol et al. [12]	Deep Learning (CNN)	Recall 96.9% Precision 34.6% F-measure 51%	403,216 PubMed citations with title and abstract as features.
Afzal et al. [15]	Support Vector Machines	ACC 92.14%	50,594 MEDLINE documents with title, abstract, publication type, and MeSH Headings as features.
Bian et al. [14]	Naïve Bayes	Recall 77.5%	15,845 PubMed citations with Scopus citation count and journal impact factor art the top two features followed by some other PubMed® metadata.
Proposed Method	Deep Learning (MLP)	ACC 97.3% AUC 0.993 Recall 95.1% F1 Score 90.4% Precision 86.25%	7958 Cochrane Review/PubMed with title and abstract as features.

Afzal et al. [15] used the Quality Recognition Model (QRM). The QRM is a supervised classification model based on the SVM machine learning algorithm trained on a dataset of clinical hedges annotated by a team of professionals [21]. This was our prior work, where we compared multiple machine learning algorithms and found SVM as the best performer. In this study, we compared our deep learning model with the best performer and other machine learning algorithms and found an increase of 5% in the accuracy. The study conducted by Bian et al. [14] used 15,845 PubMed documents and obtained 77.5% recall using the high-impact Naïve Bayes classifier. The authors found that Scopus citations and journal impact factors are two key features as compared to other features such as the number of comments on PubMed, high-impact journals, Altmetric scores, and other PubMed metadata.

4.5. Error Analysis

Cochrane Systematic Reviews are usually not up to date because of the dependency on human experts. Therefore, the training dataset lacks the inclusion of the current studies available in the primary literature, at least for the class of high-quality medical evidence. As a result, there is an imbalance between high-quality medical evidence literature and primary literature data, which may cause an error. It may contain data that has not been fully validated because the proposed model was trained on the data labeled without the help of medical professionals.

4.6. Limitation and Future Work

We have created the literature data obtained from the review of the Cochrane Library as high-quality literature data. Firstly, we need more high-quality medical evidence literature data, which is usually not available in the Cochrane Library for specific topics. Secondly, the up-to-date research papers were not included in the training model as Cochrane Library lacks the reviews for recent research. Lastly, the proposed model uses only text data such as title and abstract and is not tested by adding features such as year of publication and impact factor.

These limitations could be addressed by increasing the dataset combining data retrieved through PubMed Clinical Queries with the Cochrane Reviews data. Alternatively, the expertise of medical professionals could be utilized for the evaluation of a dataset of testing queries—add the experts' annotated data to the training dataset at a particular stage and retrain the model.

5. Conclusions

In evidence-based medicine, clinicians need high-quality medical evidence data to provide the best guidance. We proposed a deep learning model that automatically classifies biomedical literature by learning high-quality studies and general studies data. The proposed model utilized data from Cochrane Systematic Reviews and PubMed primary literature. We evaluated the proposed model based on different performance parameters like accuracy, F1-measure, and area under the curve (AUC) and obtained reasonably better results compared to competitors. We tested the model on different medical queries in the same domain of which training dataset was accumulated and also in a different domain. Experiments on four different queries in the same domain yielded similar performance to those we evaluated for training data, while the performance was slightly lower for the different domains. This research on automatic identification of high-quality evidentiary documents will reduce the work hours of clinicians who practice evidence-based medicine. It will also increase confidence of users in the decisions made by the physicians assisted with decision-making systems. In this sense, our proposed model will be a useful addition to the scientific world as well as real-world clinical setups for its vast utility in clinical practice and research.

Author Contributions: M.A. is the principal researcher who conceptualized the idea. B.J.P. did contribute towards data preparation and implementation. M.H. reviewed the text and validate the results. S.L. supervised the work and arranged funding for the project. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2017-0-01629) supervised by the IITP (Institute for Information & communications Technology Promotion)". In addition, this work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (No.2017-0-00655), NRF-2016K1A3A7A03951968, and NRF-2019R1A2C2090504This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Heneghan, C.; Mahtani, K.R.; Goldacre, B.; Godlee, F.; Macdonald, H.; Jarvies, D. Evidence based medicine manifesto for better healthcare. *BMJ* **2017**, *357*, j2973. [CrossRef] [PubMed]
2. Güiza, F.; Ramon, J.; Bruynooghe, M. Machine learning techniques to examine large patient databases. *Best Pract. Res. Clin. Anaesthesiol.* **2009**, *23*, 127–143.
3. Zhang, Y.; Lin, H.; Yang, Z.; Wang, J.; Sun, Y.; Xu, B.; Zhao, Z. Neural network-based approaches for biomedical relation classification: A review. *J. Biomed. Inform.* **2019**, *99*, 103294. [CrossRef] [PubMed]
4. Burns, G.A.; Li, X.; Peng, N. Building deep learning models for evidence classification from the open access biomedical literature. *Database* **2019**, *2019*, 1–9. [CrossRef] [PubMed]
5. McCartney, M.; Treadwell, J.; Maskrey, N.; Lehman, R. Making evidence based medicine work for individual patients. *BMJ* **2016**, *353*, i2452. [CrossRef] [PubMed]
6. Krauthammer, M.; Nenadic, G. Term identification in the biomedical literature. *J. Biomed. Inform.* **2004**, *37*, 512–526. [CrossRef] [PubMed]
7. Kilicoglu, H.; Demner-Fushman, D.; Rindflesch, T.C.; Wilczynski, N.L.; Haynes, R.B. Towards Automatic Recognition of Scientifically Rigorous Clinical Research Evidence. *J. Am. Med. Inform. Assoc.* **2009**, *16*, 25–31. [CrossRef] [PubMed]
8. Shi, H.; Liu, Y. *Naïve Bayes vs. Support. Vector Machine: Resilience to Missing Data*; Springer: Berlin/Heidelberg, Germany, 2011; pp. 680–687.
9. Khan, A.; Khan, A.; Baharudin, B.; Lee, L.H.; Khan, K.; Tronoh, U.T.P. A Review of Machine Learning Algorithms for Text-Documents Classification. *J. Adv. Inf. Technol. VOL* **2010**, *1*, 4–20.
10. Anderlucci, L.; Guastadisegni, L.; Viroli, C. Classifying textual data: Shallow, deep and ensemble methods. *arXiv* **2019**, arXiv:1902.07068.
11. Kim, Y. Convolutional neural networks for sentence classification. In Proceedings of the EMNLP 2014-2014 Conference on Empirical Methods in Natural Language Processing, Doha, Qatar, 25–29 October 2014; pp. 1746–1751.

12. Del Fiol, G.; Michelson, M.; Iorio, A.; Cotoi, C.; Haynes, R.B. A Deep Learning Method to Automatically Identify Reports of Scientifically Rigorous Clinical Research from the Biomedical Literature: Comparative Analytic Study. *J. Med. Internet Res.* **2018**, *20*, e10281. [CrossRef] [PubMed]
13. Sarker, A.; Mollá, D.; Paris, C. Automatic evidence quality prediction to support evidence-based decision making. *Artif. Intell. Med.* **2015**, *64*, 89–103. [CrossRef] [PubMed]
14. Bian, J.; Morid, M.A.; Jonnalagadda, S.; Luo, G.; Del Fiol, G. Automatic identification of high impact articles in PubMed to support clinical decision making. *J. Biomed. Inform.* **2017**, *73*, 95–103. [CrossRef] [PubMed]
15. Afzal, M.; Hussain, M.; Haynes, R.B.; Lee, S. Context-aware grading of quality evidences for evidence-based decision-making. *Health Inform. J.* **2017**, *25*, 146045821771956. [CrossRef] [PubMed]
16. Satterlee, W.G.; Eggers, R.G.; Grimes, D.A. Effective Medical Education: Insights From the Cochrane Library. *Obstet. Gynecol. Surv.* **2008**, *63*, 329–333. [CrossRef] [PubMed]
17. Armstrong, R.; Hall, B.J.; Doyle, J.; Waters, E. "Scoping the scope" of a cochrane review. *J. Public Health (Bangkok)* **2011**, *33*, 147–150. [CrossRef] [PubMed]
18. Bethesda (MD): National Center for Biotechnology Information (US). Entrez Programming Utilities Help. 2010. Available online: https://www.ncbi.nlm.nih.gov/books/NBK25501/ (accessed on 1 July 2020).
19. Winter, D.J. Rentrez: An R package for the NCBI eUtils API. *R J.* **2017**, *9*, 520–526. [CrossRef]
20. Rapidminer Build Predictive Models, Faster & Better|RapidMiner Auto Model. Available online: https://rapidminer.com/products/auto-model/ (accessed on 18 July 2020).
21. Wilczynski, N.L.; Morgan, D.; Haynes, R.B. An overview of the design and methods for retrieving high-quality studies for clinical care. *BMC Med. Inform. Decis. Mak.* **2005**, *5*, 1–8. [CrossRef] [PubMed]