Classification of Provenance Triples for Scientific Reproducibility: A Comparative Evaluation of Deep Learning Models in the ProvCaRe Project

Joshua Valdez[1], Matthew Kim[2], Michael Rueschman[2], Susan Redline[2], Satya S. Sahoo[1]

[1] Department of Population and Quantitative Health Sciences, School of Medicine, Case Western Reserve University, Cleveland OH 44106, USA
[2] Department of Medicine, Brigham and Women's Hospital and Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA
{satya.sahoo@case.edu}

Abstract. Scientific reproducibility is key to the advancement of science as researchers can build on sound and validated results to design new research studies. However, recent studies in biomedical research have highlighted key challenges in scientific reproducibility as more than 70% of researchers in a survey of more than 1500 participants were not able to reproduce results from other groups and 50% of researchers were not able to reproduce their own experiments. Provenance metadata is a key component of scientific reproducibility and as part of the Provenance for Clinical and Health Research (ProvCaRe) project, we have: (1) identified and modeled important provenance terms associated with a biomedical research study in the S3 model (formalized in the ProvCaRe ontology); (2) developed a new natural language processing (NLP) workflow to identify and extract provenance metadata from published articles describing biomedical research studies; and (3) developed the ProvCaRe knowledge repository to enable users to query and explore provenance of research studies using the S3 model. However, a key challenge in this project is the automated classification of provenance metadata extracted by the NLP workflow according to the S3 model and its subsequent querying in the ProvCaRe knowledge repository. In this paper, we describe the development and comparative evaluation of deep learning techniques for multi-class classification of structured provenance metadata extracted from biomedical literature using 12 different categories of provenance terms represented in the S3 model. We describe the application of the Long Term Short Memory (LSTM) network, which has the highest classification accuracy of 86% in our evaluation, to classify more than 48 million provenance triples in the ProvCaRe knowledge repository (available at: https://provcare.case.edu/).

Keywords: Scientific Reproducibility, Semantic Provenance, Provenance for Clinical and Health Research, Provenance Triple Classification, Deep Learning.

1 Introduction

Reproducibility is a key component of advancing scientific research that enables validation of both research protocols and study data [1, 2]. However, there is growing concern in the biomedical research domain regarding the lack of reproducible results due to missing information or lack of appropriate contextual metadata describing various aspects of a research study. For example, research study results published in peer-reviewed articles often lack details regarding the statistical models used to analyze data and the parameters used to select or discard study data for further analysis, which often leads to selection bias [2]. The lack of reproducibility has significant impact on the quality as well as integrity of published scientific results, potential misallocation of limited funding resources, and concern for patient safety during clinical trials [3]. A number of initiatives in the biomedical domain have focused on supporting scientific reproducibility, including the US National Institutes of Health (NIH) "Rigor and Reproducibility Guidelines" [4], and the Consolidated Standards of Reporting Trials (CONSORT) guidelines [5]. Provenance metadata representing essential contextual information about research studies is central to achieving the goals of the community-initiated guidelines and ensure scientific reproducibility [6].

The Provenance for Clinical and Health Research (ProvCaRe) project is developing a provenance-enabled framework to identify, characterize, and evaluate provenance metadata terms in support of scientific reproducibility [7]. The ProvCaRe project has developed: (1) the S3 model for representing multiple aspects of a research study by extending the W3C PROV Data Model (PROV-DM) [8]; (2) a provenance-focused Natural Language Processing (NLP) workflow for extracting structured provenance metadata from unstructured full-text articles from the National Center for Biotechnology Information (NCBI) PubMed [9]; and (3) the ProvCaRe knowledge repository consisting of 48.6 million provenance "triples" extracted from more than 435,000 full-text articles [7]. The ProvCaRe S3 model consists of three core concepts of:

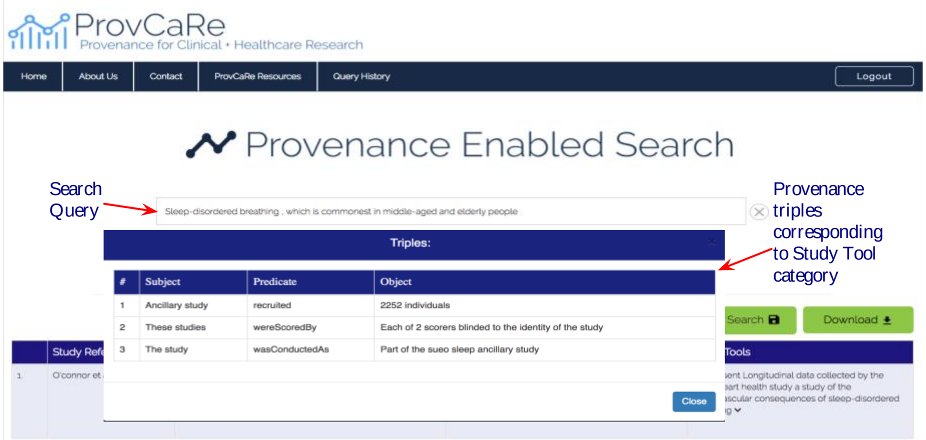
(a) **Study Method** describing the research study protocols used for data collection, inclusion-exclusion criteria among other provenance information
(b) **Study Data** describing the categorical and continuous variables used in the research study dataset, including valid data range
(c) **Study Tool** describing the hardware and software tools used for recording and analyzing research study data

Detailed provenance metadata associated with a research study, for example study design, statistical data analysis techniques, among other terms are modeled as sub-categories of these three core concepts. The S3 model has been formalized in the ProvCaRe ontology that extends the W3C PROV Ontology (PROV-O) [10] with classes and properties representing various metadata information of research studies [7]. The ProvCaRe NLP workflow uses the S3 model to identify and extract provenance metadata associated with a research study described in a full-text published article. The extracted provenance information is transformed into a triple structure similar to the W3C Resource Description Framework (RDF) model [11] with mappings to the ProvCaRe ontology terms, for example *electroencephalogram* → *wasRecordedUsing* → *scalp electrodes*. These semantic provenance triples are aggregated to create provenance graphs, which can be analyzed for characterizing the reproducibility of research studies, in the ProvCaRe knowledge repository (accessible at: https://provcare.case.edu/). The ProvCaRe knowledge repository features multiple functionalities to allow users to query and explore provenance information associated with research studies, including a hypothesis-driven search interface and a provenance-based ranking technique to rank query results. The ProvCaRe knowledgebase stores the provenance triples generated by the NLP workflow after categorizing each triple according to the S3 model concepts, which allows users to easily view provenance metadata relevant to a specific aspect of research study. For example, researchers often analyze detailed information regarding the design of the study in the context of the research hypothesis of the study, the appropriateness of the sample size of the study, and validity of the conclusions derived from the study.

Motivation and Contribution. Figure 1 shows a screenshot of provenance triples extracted from a research study exploring the association between sleep disordered breathing and hypertension, which are classified according to the S3 model. However, the classification of provenance triples according to the S3 model is a significant challenge due to multiple issues, including:

1. **Complexity** of provenance metadata information modeled in each triple requires significant effort for accurate classification;
2. **Large volume** of provenance triples generated from the ProvCaRe NLP workflow (e.g., 48 million triples generated from 435,000 papers) requires the use of automated classification techniques

The classification task for these semantic provenance triples [12] is similar to the well-known task of sentence classification in the NLP domain [13]. In particular, deep learning techniques have been used in NLP applications for classification of word vectors learned from unstructured text and have generated high quality results. Therefore, we adapted deep learning architecture used for sentence classification to classify provenance triples in the ProvCaRe project. In this paper, we describe the extension of three deep learning techniques: (1) Convoluted Neural Network (CNN); (2) Recurrent Neural Network (RNN); and (3) a combined CNN/RNN-based approach, for multi-label classification of provenance triples in the ProvCaRe project.



Sleep-disordered breathing , which is commonest in middle-aged and elderly people

#	Subject	Predicate	Object
1	Ancillary study	recruited	2252 individuals
2	These studies	wereScoredBy	Each of 2 scorers blinded to the identity of the study
3	The study	wasConductedAs	Part of the sueo sleep ancillary study

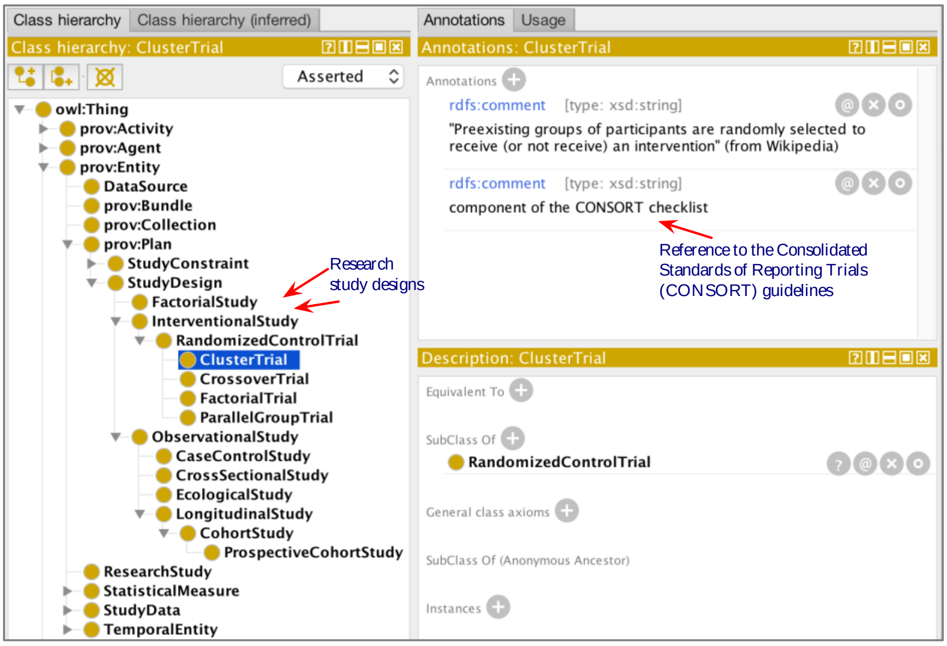
Figure 1: A screenshot of the ProvCaRe knowledgebase interface with provenance triples corresponding to a user query

2 Method

In this section, we describe the details of the S3 model used as the reference model for classification of the provenance triples, the provenance-specific training data, and details of the three deep learning models used in ProvCaRe. The deep learning models used in the ProvCaRe project were constructed using Google Tensorflow [14] and used hyperparameter tuning for classification of the provenance triples.

2.1 ProvCaRe S3 Model and Ontology

The ProvCaRe S3 model has been developed based on the NIH Rigor and Reproducibility guidelines that describe the essential components of a research study, which need to be reported in a transparent manner to support reproducibility [2]. The S3 model is modeled in a formal ontology by extending the three core classes of the W3C PROV Ontology, that is, `prov:Entity`, `prov:Activity`, and `prov:Agent` [10] (`prov` represents the W3C PROV namespace, *http://www.w3.org/ns/prov#*). The ProvCaRe ontology represents various components of the S3 model in a class hierarchy, for example three different categories of research study design, `provcare:FactorialStudy`, `provcare:InterventionalStudy`, `provcare:ObservationalStudy`, are modeled as subclasses of randomized controlled trial class (provcare represents the namespace *http://www.case.edu/ProvCaRe/provcare#*).



Annotations

rdfs:comment [type: xsd:string]
"Preexisting groups of participants are randomly selected to receive (or not receive) an intervention" (from Wikipedia)

rdfs:comment [type: xsd:string]
component of the CONSORT checklist

SubClass Of
RandomizedControlTrial

Research study designs

Reference to the Consolidated Standards of Reporting Trials (CONSORT) guidelines

Figure 2: A screenshot of the ProvCaRe ontology representing different categories of research study design.

Figure 2 shows a subset of the ProvCaRe ontology class hierarchy representing various types of research study design. Although, the ProvCaRe ontology currently models more than 1300 provenance-specific classes, it is impractical to model provenance terms for different biomedical domains using only pre-coordinated class expressions [15]. Therefore, we have developed a post-coordinated compositional grammar syntax that can be used to represent new class expressions based on requirements of specific disciplines of biomedical research [16]. Similar to the SNOMED CT post coordinated grammar syntax, this provenance-specific compositional grammar syntax allows the re-use of existing ontology terms to create new provenance expressions. For example, the expression |`Models`|: |`underwent`| = |`10-fold cross validation`| describes the validation method for a model used in a prospective cohort study to evaluate association between sleep disordered breathing and hypertension [17]. Together with pre-coordinated classes, the post coordinated syntax enables the representation of a variety of provenance terms in the ProvCaRe ontology.

The ProvCaRe ontology is used as the reference knowledge model in the NLP workflow for named entity recognition (NER), generation of provenance triples from parse tree, and finally classification of the provenance triples before they are added to the ProvCaRe knowledgebase for user query. The classification of provenance triples using deep learning model enables easier visualization of query results for users.

2.2 Training of Deep Learning Models

We used two manually created datasets consisting of provenance metadata extracted from 75 full-text articles describing sleep medicine research studies to train the deep learning models. The 75 articles were selected by two members of the US NIH-funded National Sleep Research Resource (NSRR) project, which is creating the largest repository of sleep medicine study data from more than 40,000 sleep studies involving more than 36,000 participants [18]. They selected published articles describing research studies that are releasing their study data through the NSRR project. Therefore, the provenance information extracted from these articles can be used by the sleep medicine community to potentially reproduce the results reported in these published articles. These manually extracted provenance triples serve as gold standard in the ProvCaRe project and they are used to train the deep learning models.

As part of the training procedure, the first step involves defining a session and creation of a default Tensorflow graph in that session (a graph in Tensorflow can be understood as a structure that contains tensors and operations). The next step in the training process defines the Tensorflow network's loss function and optimizes the loss function using built-in Tensorflow Adam optimizer. Once the optimizer has been defined, a function is created to perform a single training step using the optimizer. This is implemented by allowing Tensorflow to automatically detect variables that can be trained and then calculate the gradients of these variables. In the next phase, the global step is defined and passed to the optimizer. This allows the count of the training steps to be computed automatically by Tensorflow. The final phase in the training process involves looping through the training steps defined above using specific number of training loops that iterate over the predefined batched data.

2.3 Deep Learning Model Architectures

In this section, we describe the details of the three deep learning models used to classify the provenance triples in the ProvCaRe project.

Convoluted Neural Network (CNN). The CNN model used for classification of provenance triples is similar to a CNN architecture proposed by Kim et al. [13]. This CNN model has been used in the NLP community and it has performed well on a variety of tasks ranging from sentiment analysis to classification. The architecture for this model comprises of a first layer that embeds each word of the sentence into a low-dimensional vector space using the pre-trained word-2-vec vectors. Following word embedding, the vectors are passed to a convolution layer which performs convolutions over the embedded vector using the specified filter sizes (standard filter sizes of 3,4 and 5 were used in this project). Once the convolutions have been performed, the results are max-pooled into a single, large feature vector. In the next step, dropout regularization is added to stochastically turn-off a portion of the model to force the network to learn features individually and not together (drop out is only set during the training process and is disabled during prediction). Finally, the result is classified with a softmax layer and a prediction label is produced with the highest value by performing matrix multiplication operation. Figure 3 illustrates the components of a CNN model used in this project.

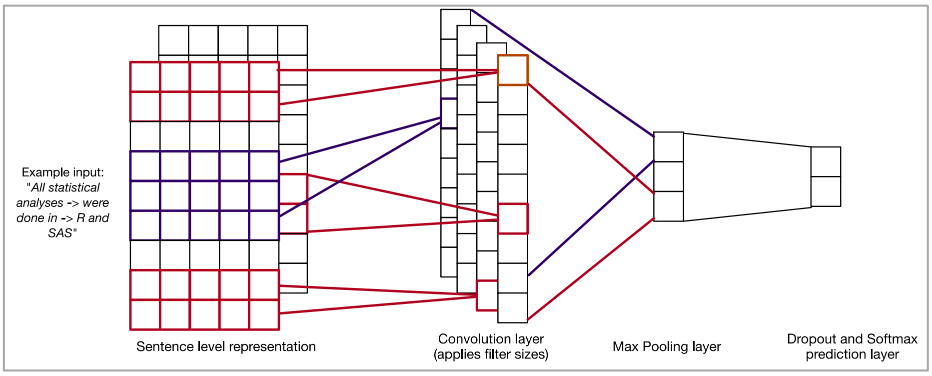


Figure 3: Overview of the CNN model used in the ProvCaRe project

Recurrent Neural Network (RNN). RNN is the second deep learning model evaluated in the project for classification of the provenance triples. Like CNN's, RNN's are a type of neural network which have become more popular in recent years for natural language classification. Specifically, RNN's feedforward networks consist of recurrent connections. The advantage of this approach is that these connections afford the network the capacity to refer to previous states. This means that RNN's are able to process arbitrary sequences of input. RNN's model define and build a language model. This is implemented through a series steps. The first step in this process is to clean and store the training data. This is done according to several NLP techniques. To accomplish this in our model we tokenize the text, remove stop words as well as infrequent words and add special tokens to the beginning and end of the sentence. Figure 4 illustrates an overview of a RNN network.

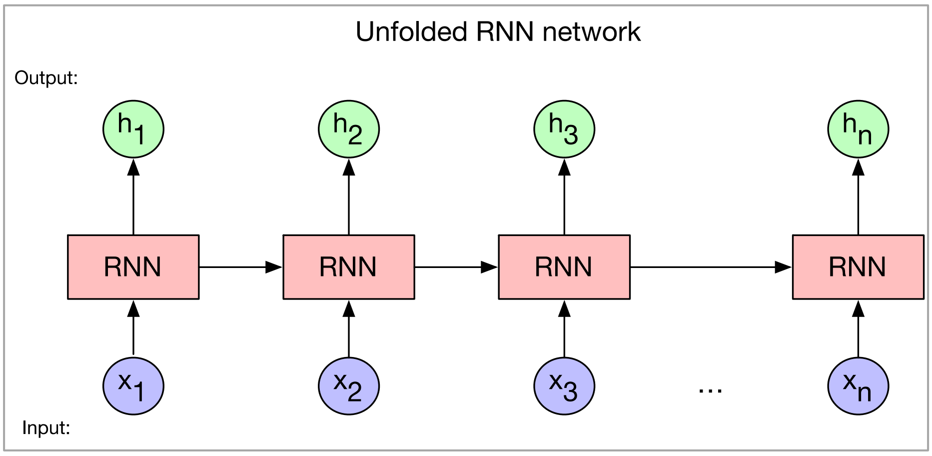


Figure 4: Overview of the RNN model with input and output variables

Once the training data has been cleaned, the next step is to construct the data matrices that will hold our training data and map the words in the sentence to a vector representation. Once our vectors have been built on our training data, the next step is to feed our data into the RNN model. Our model represents each word from the training data as a "one-hot" vector. Once the data has been converted a Tensorflow RNN, it is initialized with the parameters specified above. During initialization, we allow for forward propagation, this returns both the calculated outputs and the hidden states which are used to calculate the gradients using Backpropagation Through Time (BPTT). Once this is done full the full training procedure is performed making use of stochastic gradient descent and BPTT the predicted label is produced.

Long-Short Term Memory (LSTM or CNN/RNN). Given the wide acceptance of neural network architectures in the NLP community, there has been recent interest in the use of a combined approach for classification tasks. To this end, we implemented a combined model consisting of our CNN and RNN approaches described above. This model is created with pre-trained vectors from word2vec, max pooling and LSTM recurrent unit. The model takes local features extracted by the CNN as input for the RNN. This is accomplished by first using the predefined word embedding as the input for the CNN. The output of this step are feature maps, which were described earlier as part of the CNN implementation and are formed during the convolutional windowing process. After convolution and pooling operations, the encoded feature maps are fed as input into the RNN. The output from the RNN are the learned sentence-level representations which are given to the network and the softmax output which produces the classification label.

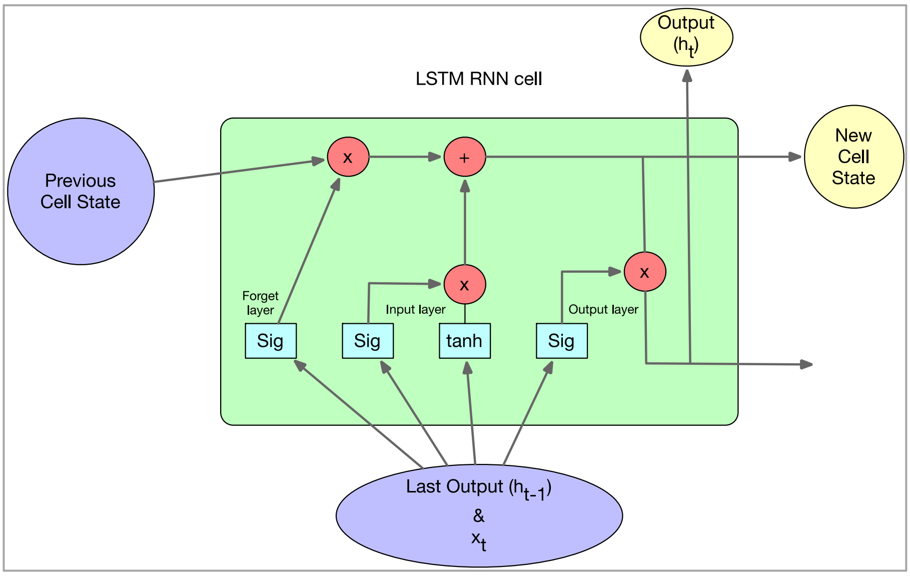


Figure 5: Overview of the LSTM model with input, output, and other layers of the network

These three deep learning models were applied to classify the provenance triples generated by the ProvCaRe NLP workflow and the classified triples were added to the ProvCaRe knowledgebase.

3 Result and Discussion

In this section, we describe the results of a comparative evaluation of the three deep learning models for classification of provenance triples.

3.1 Classification Results

Table 1 shows the training and test accuracy of the three deep learning models in addition to the time taken to train the three models. The training and test results are evaluated using the following hyperparameter values: number of epochs (1); batch size (50); number of filters (32); filter sizes (3, 4, 5, 7); embedding dimension (200); evaluation loop (100); hidden unit (300); I1 regularization lamda (0); and dropout keep probability (0.5). The CNN and LSTM models have comparable accuracy in terms of training with score of 0.904 and 0.909 respectively. Similarly, the test accuracy for CNN and LSTM are 0.844 and 0.861 respectively. In contrast, the RNN model has comparatively low accuracy of 0.792 for training and 0.758 for test. The test accuracy is evaluated using a manually annotated dataset of provenance triples created by the two members of the NSRR project serving as domain experts. It is important to note that the training and test accuracy for all three deep learning models are similar demonstrating the effectiveness of the training process.

Based on the results of the comparative evaluation, we integrated the LSTM model for classification of provenance triples in the ProvCaRe project. In the next section, we describe the results of classifying five datasets of provenance triples generated by the ProvCaRe NLP workflow.

Table 1: Comparative evaluation of the test and training accuracy of the three deep learning models.

	CNN	RNN	CNN/RNN (LSTM)
Training Accuracy	0.904332007	0.792777778	0.909607732
Test Accuracy	0.844146341	0.758648649	0.86105042
Training Time	14 minutes 47 seconds	13 minutes 05 seconds	16 minutes 23 seconds

3.2 Comparative Evaluation Results

The LSTM deep learning model was used to evaluate 5 datasets of provenance triples consisting of: (1) 20,000 triples; (2) 50,000 triples; (3) 100,000 triples; (4) 500,000 triples; and (5) 1 million triples, which were added to the ProvCaRe knowledgebase. To systematically evaluate the multi-label classification feature of the deep learning models, we used 12 subcategories of the S3 model. Table 2 shows the distribution of provenance triples across the 12 subcategories of the S3 model. The results show that Data Collection, Study Hypothesis, and terms describing Comparison (a component of the PICO(T) model used in evidence-based medicine [19]) are the three categories with highest number of provenance triples across all the five datasets. In contrast, the S3 subcategory describing software tools used in a research study have the lowest number of provenance triples. This is not surprising as only relatively few number of research studies use software tools, such as R or SAS-based libraries for data pre-processing or data analysis, therefore the occurrence of provenance triples describing software tools is relatively low.

Table 2: Results of classification of provenance triples according to the S3 model.

Triples	**SV**	**SH**	**SM**	**S**	**P**	**RI**	**I**	**DP**	**DC**	**DA**	**C**	**TV**
20000	798	3,154	1,861	39	1,572	2,233	904	425	3,157	1,819	3,821	217
50000	2926	7,094	4,645	655	4,728	4,219	3,077	1,857	6,787	3,998	9,213	801
100000	4652	12,003	10,295	903	10,243	7,428	7,164	2,914	14,399	8,986	18,896	2,117
500000	20876	61,399	49,488	5,502	52,339	37,224	31,359	16,898	70,853	42,597	98,755	12,710
1000000	39452	127,125	95,820	9,039	100,409	80,003	58,862	36,039	146,319	87,343	197,713	21,876

Table legends: SV (Study Variables), SH (Study Hypothesis), SM (Statistical Model), S (Software), P (Population), RI (Recording Instrument), I (Intervention), DP (Data Preprocessing), DC (Data Collection), DA (Data Analysis), C (Comparison), TV (Time Value)

It is interesting to note that a high number of provenance triples are categorized into the important S3 subcategories of recording instruments and statistical models with 80,003 and 95, 820 triples respectively. Provenance metadata describing the details of the recording instruments used in a research study is important to provide essential contextual information for interpreting the research study data. For example, provenance information describing the specific type of blood pressure measurement instrument used to record systolic and diastolic blood pressure of patients and scalp or intracranial electrodes used to record brain electrical activity are important for subsequent analysis of the recorded data. Similarly, a high number of provenance triples describe the details of research study population (100,409 provenance triples out of 1 million provenance triples), which is critical for evaluating the statistical significance of the results reported in a given study.

In contrast, the provenance triples describing the data variables used in a research study and time values associated with different aspects of a research study (e.g., recording of patient information) is relatively low. This may highlight the need for encouraging researchers to improve the quality of provenance metadata describing these two important aspects of a research study. Overall, this analysis of the distribution of provenance triples according to the S3 model subcategories clearly highlights the need for classification of provenance triples for analyzing the properties of provenance metadata extracted from published biomedical articles. In addition to the ease of query result visualization (as illustrated in Figure 1), the classification of provenance triples using deep learning model (as described in this paper) makes it easier to characterize the provenance metadata available in published articles describing biomedical research studies.

4 Conclusion

The ProvCaRe project aims to advance the use of provenance metadata to meet the objectives of various community-based initiatives to improve scientific reproducibility in the biomedical research domain. In this paper, we described a comparative evaluation of deep learning models to address the critical challenge of automated and accurate classification of semantic provenance triples generated by the ProvCaRe NLP workflow. The three deep learning models were trained and evaluated using a manually curated dataset of provenance triples generated from 75 papers describing sleep medicine research studies (as part of the NSRR project).

The LSTM model featuring a combination of CNN and RNN outperformed both CNN and RNN models individually. The LSTM model was used to classify five dataset of provenance triples according to 12 subcategories of the S3 model, which were analyzed to demonstrate the importance of provenance triple classification for easier analysis and interpretation of provenance metadata extracted from published biomedical articles.

Acknowledgement. This work is supported in part by the NIH-NIBIB Big Data to Knowledge (BD2K) 1U01EB020955 grant, NSF grant#1636850, and the NIH-NHLBI R24HL114473 grant.

References

[1] F. S. Collins, Tabak, L.A., "Policy: NIH plans to enhance reproducibility," *Nature*, vol. 505, pp. 612-613, 2014.

[2] S. C. Landis, Amara, S.G., Asadullah, K., Austin, C.P., Blumenstein, R., Bradley, E.W., Crystal, R.G., Darnell, R.B., Ferrante, R.J., Fillit, H., Finkelstein, R., Fisher, M., Gendelman, H.E., Golub, R.M., Goudreau, J.L., Gross, R.A., Gubitz, A.K., Hesterlee, S.E., Howells, D.W., Huguenard, J., Kelner, K., Koroshetz, W., Krainc, D., Lazic, S.E., Levine, M.S., Macleod, M.R., McCall, J.M., Moxley, R.T. 3rd, Narasimhan, K., Noble, L.J., Perrin, S., Porter, J.D., Steward, O., Unger, E., Utz, U., Silberberg, S.D., "A call for transparent reporting to optimize the predictive value of preclinical research," *Nature,* vol. 490, no. 7419, pp. 187-91, 2012.

[3] F. Prinz, Schlange, T., Asadullah, K., "Believe it or not: how much can we rely on published data on potential drug targets?," *Nature Reviews Drug Discovery,* vol. 10, no. 9, p. 712, 2011.

[4] National Institutes of Health. (2016). *Principles and Guidelines for Reporting Preclinical Research*. Available: https://www.nih.gov/research-training/rigor-reproducibility/principles-guidelines-reporting-preclinical-research

[5] K. F. Schulz, Altman, D.G., Moher, D., CONSORT Group., "CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials.," *Journal of Clinical Epidemiology,* vol. 63, no. 8, pp. 834-40, 2010.

[6] S. S. Sahoo, Valdez, J., Rueschman, M., "Scientific Reproducibility in Biomedical Research: Provenance Metadata Ontology for Semantic Annotation of Study Description," in *American Medical Informatics Association (AMIA) Annual Symposium*, Chicago, 2016, pp. 1070-1079.

[7] J. Valdez, Kim, M., Rueschman, M., Socrates, V., Redline, S., Sahoo, S.S., "ProvCaRe Semantic Provenance Knowledgebase: Evaluating Scientific Reproducibility of Research Studies " presented at the American Medical Informatics Association (AMIA) Annual Conference, Washington DC, 2017.

[8] L. Moreau, Missier, P., "PROV Data Model (PROV-DM)," in "W3C Recommendation," World Wide Web Consortium W3C2013.

[9] J. Valdez, Rueschman, M., Kim, M., Redline, S., Sahoo, S.S., "An Ontology-Enabled Natural Language Processing Pipeline for Provenance Metadata Extraction from Biomedical Text.," presented at the 15th International Conference on Ontologies, DataBases, and Applications of Semantics (ODBASE) 2016.

[10] T. Lebo, Sahoo, S.S., McGuinness, D., "PROV-O: The PROV Ontology," in "W3C Recommendation," World Wide Web Consortium W3C2013.

[11] I. Herman, Adida, B., Sporny, M., Birbeck, M., "RDFa 1.1 Primer - Second Edition," in "W3C Working Group Note," World Wide Web Consortium (W3C)2013, Available: http://www.w3.org/TR/rdfa-primer/.

[12] S. S. Sahoo, Sheth, A., Henson, C., "Semantic Provenance for eScience: Managing the Deluge of Scientific Data," *IEEE Internet Computing,* vol. 12, no. 4, pp. 46-54, 2008.

[13] Y. Kim, "Convolutional neural networks for sentence classification," *arXiv preprint* Available: https://arxiv.org/abs/1408.5882

[14] *TensorFlow*. Available: https://www.tensorflow.org/

[15] A. L. Rector, Brandt, S., Schneider, T., "Getting the foot out of the pelvis: modeling problems affecting use of SNOMED CT hierarchies in practical applications," *J Am Med Inform Assoc.*, vol. 18, no. 4, pp. 432–440, 2011.

[16] J. Valdez, Rueschman, M., Kim, M., Arabyarmohammadi, S., Redline, S., Sahoo, S.S., "An Extensible Ontology Modeling Approach Using Post Coordinated Expressions for Semantic Provenance in Biomedical Research," in *The 16th International Conference on. Ontologies, DataBases, and Applications of Semantics (ODBASE)*, Rhodes, Greece, 2017.

[17] G. T. O'Connor, Caffo, B., Newman, A.B., Quan, S.F., Rapoport, D.M., Redline, S., Resnick, H.E., Samet, J., Shahar, E., "Prospective study of sleep-disordered breathing and hypertension: the Sleep Heart Health Study," *Am J Respir Crit Care Med*, vol. 179, no. 12, pp. 1159-64, 2009.

[18] D. A. Dean, Goldberger, A.L., Mueller, R., Kim, M., Rueschman, M., Mobley, D., Sahoo, S.S., Jayapandian, C.P., Cui, L., Morrical, M.G., Surovec, S., Zhang, G.Q., Redline, S., "Scaling up scientific discovery in sleep medicine: the National Sleep Research Resource.," *SLEEP*, vol. 39, no. 5, pp. 1151–1164, 2016.

[19] X. Huang, Lin, J., Demner-Fushman, D., "Evaluation of PICO as a knowledge representation for clinical questions," presented at the AMIA Annual Symposium Proceedings, 2006.