MMLN: Leveraging Domain Knowledge for Multimodal Diagnosis

Haodi Zhang
Shenzhen University
College of Computer and Software Engineering
Shenzhen, China
hdzhang@szu.edu.cn

Chenyu Xu
Shenzhen University
College of Computer and Software Engineering
Second People’s Hospital of Guangdong Province
Shenzhen, China
2060271017@email.szu.edu.cn

Peirou Liang
Shenzhen University
College of Computer and Software Engineering
Shenzhen, China
2020155016@email.szu.edu.cn

Ke Duan
Shenzhen University
College of Computer and Software Engineering
Shenzhen, China
2019151087@email.szu.edu.cn

Hao Ren
Second People’s Hospital of Guangdong Province
Guangzhou, China
1810272037@email.szu.edu.cn

Weibin Cheng
Second People’s Hospital of Guangdong Province
Guangzhou, China
chwb817@gmail.com

Kaishun Wu
Shenzhen University
College of Computer and Software Engineering
Shenzhen, China
wu@szu.edu.cn

ABSTRACT
Recent studies show that deep learning models achieve good performance on medical imaging tasks such as diagnosis prediction. Among the models, multimodality has been an emerging trend, integrating different forms of data such as chest X-ray (CXR) images and electronic medical records (EMRs). However, most existing methods incorporate them in a model-free manner, which lacks theoretical support and ignores the intrinsic relations between different data sources. To address this problem, we propose a knowledge-driven and data-driven framework for lung disease diagnosis. By incorporating domain knowledge, machine learning models can reduce the dependence on labeled data and improve interpretability. We formulate diagnosis rules according to authoritative clinical medicine guidelines and learn the weights of rules from text data. Finally, a multimodal fusion consisting of text and image data is designed to infer the marginal probability of lung disease. We conduct experiments on a real-world dataset collected from a hospital. The results show that the proposed method outperforms the state-of-the-art multimodal baselines in terms of accuracy and interpretability.

CCS CONCEPTS
• Applied computing → Health informatics; • Computing methodologies → Mixture models.

KEYWORDS
domain knowledge, multimodal, Markov logic network, disease diagnosis

ACM Reference Format:
Haodi Zhang, Chenyu Xu, Peirou Liang, Ke Duan, Hao Ren, Weibin Cheng, and Kaishun Wu. 2018. MMLN: Leveraging Domain Knowledge for Multimodal Diagnosis. In Woodstock ’18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY. ACM, New York, NY, USA, 8 pages.
https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION
In recent years, advances in deep learning and the release of multiple, large, publicly available chest X-ray (CXR) datasets have led to a promising performance in many medical imaging analysis tasks. Among all medical imaging approaches, CXR is the most common medical radiological examination in the world. As a particularly important modality, many lung diseases are first diagnosed by CXR, and a variety of CXR applications have been researched and even boosted numbers of products.

However, CXR alone is not enough. Multimodality has been an emerging trend, integrating different forms of data such as text data and image data. Since the correlation between text modality and image modality is relatively large, diagnosis accuracy...
We propose a data-driven and knowledge-driven framework called multimodal Markov logic network, which bridges domain knowledge and deep learning to help lung disease diagnosis.

We conduct experiments on a real-world dataset collected from a hospital to demonstrate the effectiveness of our model compared with the state-of-the-art baselines.

2 RELATED WORK

The task of disease diagnosis has a long history. Since the last century, people have been exploring how to use computers to assist in disease diagnosis. MYCIN [25] is one of the first expert system that used rule-based reasoning to diagnose blood infections. Incorporating about 500 production rules and certainty factors, MYCIN performed at roughly the same level of experts. Besides, people start to research Clinical decision support system (CDSS). Kunhiman et al. proposed a CDSS for diagnosis of peripheral neuropathy using fuzzy logic. Through 24 input fields which include symptoms and diagnostic test outputs, they achieved 93% accuracy compared to experts at identifying motor, sensory, mixed neuropathies, or normal cases. DXplain [18] is an electronic reference based CDSS that provides probable diagnosis based on clinical manifestations. In a randomized control trial involving 87 family medicine residents, those randomized to use the system showed significantly higher accuracy (84% vs. 74%) on a validated diagnosis test involving 30 clinical cases. The problem of pure symbolic models is the performance is limited, and the migration ability is relatively weak.

In recent years, deep learning has made a tremendous impact in the field of medical imaging, which can greatly enhance the capabilities of Image-level Prediction, Segmentation, Image Generation, Domain Adaptation, Localization, etc. With the release of several large publicly CXR datasets, including CheXpert [11], chestXray8 [27], PadChest [2], etc., the development of CXR diagnosis is accelerating. A series of state-of-the-art CNNs for image-level classification are proposed. Among the pathology, pneumonia is one of the most studied subject. Singh et al [23] compare the performance of different architectures with various depths on a given task. Sirazitdinov et al [24] evaluate the effect of various data augmentation and input pre-processing methods. During the COVID-19, Wang et al. [26] proposed a deep learning pipeline for the diagnosis and discrimination of viral, non-viral and COVID-19 pneumonia from chest X-ray images with AUC of above 0.87. [9] developed an explainable convolutional neural network based on CXR images classification and analysis for COVID-19 pneumonia detection. It can select instances of CXR images to explain the behavior to achieve higher prediction accuracy of above 96%. Although pure machine learning based methods have a good performance, they only learn from data and output the probability. They all belong to probabilistic models, lack of the support of knowledge.

There are some work incorporate knowledge graph (KG) or ontology as supplements to knowledge. For example, [17] propose a 8-step pipeline to build a medical knowledge graph (KG) with a novel quadruplet structure from EMRs. Then the KG is used to...
3 PROPOSED METHOD

We propose a simple yet effective multimodal Markov Logic Network that effectively integrates text data and image data (Figure 2). Text data mainly includes EMRs, laboratory items and radiology reports, while image data mainly refer to CXR images. Undoubtedly, multimodal data can improve the accuracy of disease diagnosis. Chest imaging examination is an indispensable part of doctor’s diagnosis. Common lung diseases are first diagnosed using CXR. After formulating the diagnostic rules, our goal is to extract each possible grounding of each predicate appearing in rules to build evidence database for weight learning and inference. The detailed process is shown in Algorithm 1.

3.1 Formulate diagnostic rules

Figure 3 shows the clinical diagnostic criteria for community-acquired pneumonia (CAP). We use first-order logic (FOL) to represent knowledge about pneumonia clinical guidelines. For the first item, we define a query predicate called Pneumonia representing CAP. For the second item, we define evidence predicates for each clinical manifestation. For example, we define a predicate called Fever corresponding to 2 (2). For the third item, we define evidence predicates about chest imaging examination. For example, we define predicates called Infiltration, Consolidation, etc. We then use the defined predicates to construct the FOL formulas according to the logical relationship described in the last paragraph of the guideline. For the first item, all the selected pneumonia cases are CAP, so it is satisfied by default. For the second and third items, we define one of the rules as follow:

\[ w \cdot \text{Fever}(x) \land \text{Consolidation}(x) \implies \text{Pneumonia}(x) \quad (1) \]

However, the real world is complex and uncertain, especially in the domain of disease diagnosis. Therefore, we attach a weight \( w \) to each rule, which becomes a MLN. When a world violates one rule, it becomes less probable but not impossible.

3.2 Building multimodal evidence database

After formulating the diagnostic rules, our goal is to extract each possible grounding of each predicate appearing in rules to build evidence database for weight learning and inference. The detailed process is shown in Algorithm 1.

Algorithm 1 Building evidence database

Input: EMRs, CXR, Guideline
Output: Evidences

1. Begin formulate diagnostic rules
2. Define the predicates from the Guideline \( \rightarrow \) Predicates
3. Formulate rules using defined Predicates according to the logic relations of Guideline \( \rightarrow \) Rules
4. End formulate diagnostic rules
5. Finetune a CNN model to extract pathologies from CXRs
6. Begin building evidence database
7. Extract mentioned predicates about clinical manifestations from EMRs \( \rightarrow \) Evidences_EMRs
8. Extract mentioned predicates about chest imaging examination from CXRs \( \rightarrow \) Evidences_CXR
9. Combine Evidences_EMRs and Evidences_CXR to build Evidences
10. End building evidence database
positive pathology: CXR_Lung_inflammation. Therefore, we get a grounding predicate as follows:
CXR_Lung_inflammation(P450945).

Combining the grounding predicates extracted from the text data and image data, we get an evidence database, which is a large scale of observations from real-world multimodal data.

3.3 Learning weights and inference
After building the evidence database, we split the database into training set and test set in a 80/20 ratio. Training set is used to learn weights, while test set is used to infer. The detailed algorithm is shown in Algorithm 2.

### Algorithm 2: Weights learning and inference of MMLN

**Input:** Predicates, Rules, Evidences  
**Output:** Pneumonia probability

1. Initialize Nodes for each possible grounding of each predicate in Predicates
2. Add one feature for each possible grounding of each rule in Rules
3. Split the Evidences into training set Evidences_train and test set Evidences_test
4. Begin learning weights
5. Initialize the weights of Rules by zero
6. Learnwts(Predicates, Rules, Evidences_train) → Weights
7. End learning weights
8. Begin inference
9. for case_i in Evidences_test do
10. Infer(Predicates, Rules, Weights, Evidences_i) → Pneumonia_probability
11. end for
12. End inference

#### 3.3.1 Learning weights
We use the MLN tool Alchemy [14] to learn each weight of rules and inference. Given defined predicates, unweighted rules and the training set evidence database, we use function learnwts to learn each weight of rules. For example, we learn each weight of rule as follows:

**Higher weight represents more probable rule.**

#### 3.3.2 Inference
The goal of inference is to calculate the marginal probability of specific illness to make a disease diagnosis. Given
defined predicates, weighted rules and the test set of evidence
database, we use function infer to perform inference. As for the
example mentioned in section 3.2, we get the result as follows:
Pneumonia(P450945) 0.99895, the probability represents the illness probability, it shows that
the case is highly pneumonia.

4 EXPERIMENT
4.1 Dataset
We collected the multimodal data of the admitted patients from the
EMR system and Picture Archiving and Communication System
(PACS) of the hospital information department. The text data con-
sists of EMRs, laboratory items and radiology reports. The image
data mainly includes CXR images. The de-identified data collected
in this experiment all come from real medical data from the Sec-
ond People’s Hospital of Guangdong Province, and the privacy of
patients has been removed.

4.1.1 Data processing. We first find out all cases with both EMR
and CXR. In our experiment, disease diagnosis is regard as a bi-
nary classification task of pneumonia and other diseases related
to pneumonia. We treat pneumonia cases as positive samples, and
other lung diseases as negative samples. Since doctors usually give
a series of diagnostic codes rather than a single disease, we need to
screen the cases. For pneumonia cases, we select cases diagnosed
with pneumonia in admission diagnosis to ensure that they were
CAP cases. We then further screened the cases with the main diag-
nosis of pneumonia from the discharge diagnosis and preliminary
diagnosis. For negative samples, it is also necessary to additionally
exclude cases with pneumonia in the diagnosis to ensure that they
does not suffer from pneumonia.

4.2 Comparison of unimodal model and
multimodal model
In this section, we compare the performance between unimodal
and multimodal models. Unimodal means that we build evidence
database purely on text data. Then we use the unimodal database
to learn weights and inference.

As for the clinical manifestations, we adopt the same methods
mentioned in section 3.2 to extract grounding predicates. As for
the chest imaging examination, we extract grounding predicate
Lung_inflammation from EMRs instead of CXR images. We use

Figure 6: ROC of unimodal model and multimodal model.

regular expressions to match related description from EMRs’ fields
such as physical examination, auxiliary examination, case char-
acteristics, diagnosis, diagnostic basis and differential diagnosis,
general situation, specialty situation. Multimodal model is imple-
mented in multimodal disease diagnosis framework to effectively
incorporate text data and image data. A multi-label classification
pretrained model on TorchXrayVision [5] is used to output common
lung pathologies.

After learning weights and inference, we get the unimodal and
multimodal inference accuracy respectively. The performance is
shown in Table 1 and ROC is compared in Figure 6. As shown in
Table 1, the overall performances of multimodal model are better
than those of the unimodal model. In particular, our unimodal
model reaches an accuracy above 0.90. Incorporating image data,
our multimodal model has a large improvement in accuracy, which
is near to 0.97. We also draw the ROC curves for the classification
results of unimodal and multimodal model, as shown in Figure 6.
We can see that the curve of multimodal is closest to the left top
corner, and its corresponding AUC score is 0.965 which is also the
highest AUC values. Generally, the results prove that multimodal
data can actually improve the accuracy of disease diagnosis.
Table 1: Performance of our model against competitive methods

| Model                        | Accuracy | AUC    | F1     | Precision | Recall |
|------------------------------|----------|--------|--------|-----------|--------|
| MMBT(Bert-base-uncased)      | 0.897    | 0.883  | 0.916  | 0.873     | 0.963  |
| MMBT(Bert-large-uncased)     | 0.950    | 0.946  | 0.958  | 0.946     | 0.970  |
| ConcatBert                   | 0.936    | 0.929  | 0.946  | 0.924     | 0.970  |
| Ours(Unimodal)               | 0.905    | 0.948  | 0.929  | 0.927     | 0.931  |
| Ours(Multimodal)             | 0.965    | 0.983  | 0.976  | 1.000     | 0.954  |
| Ours(Multimodal, small size) | 0.924    | 0.967  | 0.948  | 0.987     | 0.911  |

4.3 Comparison of MMLN and multimodal baselines

We compare three multimodal baselines with MMLN in this experiment. MultiModal BiTransformers (MMBT) [13]: the underlying pretrained models are BERT [6] and ResNet [8]. The text data is passed through BERT to generate word embedding, while the image data is passed through ResNet to generate CNN embedding, and then the two embeddings are concatenated and passed through a classifier for disease classification. We use two models with different size to compare. ConcatBert [1]: it is another model for multimodal classification with text and images, whose text representation obtained from pretrained BERT base model and image representation obtained from VGG16 pretrained model.

For the text data, we extracted the main complaint, current medical history, physical signs, laboratory items, lung pathologies from EMRs. We concatenate these fields of text and translate it into English as text data input. For the image data, since multiple CXR images are generated as one patient has done CXR examinations by multiple times or multiple angles, we just randomly select one of the CXR images as the image data input. We select cases with both EMR and CXR, including a total of about 2000 cases. The performance is shown in Table 1.

As shown in Table 1, the accuracy and F1 score of our multimodal model are better than those of the baseline models. Specifically, three classical deep learning models achieve accuracies in the range of 0.89 to 0.95. At the same time, except MMBT with bert-base-uncased, the baselines also yield accuracies with values from 0.93 to 0.95. In comparison, our multimodal reaches an accuracy near to 0.97. The same trend can also be observed in the comparison results of F1 scores. Generally, the performance of our multimodal model is superior to those of the other competing methods based on deep learning.

4.4 MMLN robustness analysis

When we collect data in the hospital, we find that it is quite difficult to obtain multimodal data as there is a widespread problem of missing modalities. Therefore, we explore the effect of training set size on MMLN here to illustrate the model is robust on dataset size.

Our original size of cases, including pneumonia cases and other pneumonia differential diagnosis cases, is about 2400, which is a medium-sized dataset. In this experiment, we only take 1/10 of the number of each disease case, and performed weight learning and inference on a small size of dataset.

In Table 1, we can see that weights learning by large size dataset are bigger than those in small size. Table 1 lists the results of large size and small size multimodal model performance. Even if the size is only 1/10 of the original, the performance will not drop too much, the accuracy reaches 0.924, and the F1 score reaches 0.948, which is even better than MMBT (bert-base-uncased). In Figure 7, two ROC curves are very close. In general, MMLN has little effect on the size of the training data, and even if the size is small, we can get a not too bad performance.

5 CONCLUSION

In this paper, we propose a data-driven and knowledge-driven MMLN model for lung disease diagnosis. We formulate diagnosis rules in form of Markov logic according to authoritative clinical medical guidelines, which improves the accuracy and interpretability of the model. Moreover, MMLN effectively fuse multimodal data and reveal intrinsic relations between different data sources by leveraging the medical domain knowledge. The experimental results show that knowledge and data is fully complementary to better the downstream diagnosis task and MMLN outperforms the state-of-the-art multimodal models for disease classification.
6 ACKNOWLEDGMENTS

ACKNOWLEDGMENTS

REFERENCES

[1] John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A González. 2017. Gated multimodal units for information fusion. arXiv preprint arXiv:1702.01992 (2017).

[2] Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria de la Iglesia-Vayá. 2017. A simple algorithm for identifying negated findings and diseases in discharge summaries. Journal of biomedical informatics 34, 5 (2001), 301–310.

[3] Joseph Paul Cohen, Joseph D. Viviano, Paul Bertin, Paul Morrison, Parra Torabian, Matteo Guarinello, Matthew P Juntgen, Akshay Chaudhari, Rupert Brooks, Mohammad Hashir, and Hadrien Bertrand. 2020. TorchXRayVision: A library of chest X-ray datasets and models. https://github.com/mlmed/torchxrayvision (2020). https://github.com/mlmed/torchxrayvision (2020).

[4] Jingchi Jiang, Xueli Li, Chao Zhao, Yi Guan, and Qiubin Yu. 2017. Learning a health knowledge graph from electronic medical records. Nature communications 8 (2017), 1–15.

[5] Youli Fang, Hong Wang, Lutong Wang, Ruitong Di, and Yongqiang Song. 2019. Artificial intelligence medical knowledge graph: construction and applications. Future Generation Computer Systems 102 (2020), 534–548.

[6] Ramandeep Singh, Mannandeep K Kalra, Chayanin Nitiwarangkul, John A Patti, Fatemeh Homayounieh, Atul Padole, Pooja Rao, Preetham Putha, Victorine V Muse, Amita Sharma, et al. 2018. Deep learning in chest radiography: detection of findings and presence of change. PloS one 13, 10 (2018), e0204155.

[7] Ying Shen, Lixiu Zhang, Jin Zhang, Min Yang, Buzhou Tang, Yaliang Li, and Kai Lei. 2018. CNN: Constructing a clinical Bayesain network based on data from the electronic medical record. Journal of biomedical informatics 88 (2018), 1–10.

[8] Ying Shen, Lixiu Zhang, Jin Zhang, Min Yang, Buzhou Tang, Yaliang Li, and Kai Lei. 2018. CNN: Constructing a clinical Bayesain network based on data from the electronic medical record. Journal of biomedical informatics 88 (2018), 1–10.

[9] Ying Shen, Lixiu Zhang, Jin Zhang, Min Yang, Buzhou Tang, Yaliang Li, and Kai Lei. 2018. CNN: Constructing a clinical Bayesain network based on data from the electronic medical record. Journal of biomedical informatics 88 (2018), 1–10.

[10] Ying Shen, Lixiu Zhang, Jin Zhang, Min Yang, Buzhou Tang, Yaliang Li, and Kai Lei. 2018. CNN: Constructing a clinical Bayesain network based on data from the electronic medical record. Journal of biomedical informatics 88 (2018), 1–10.

[11] Ying Shen, Lixiu Zhang, Jin Zhang, Min Yang, Buzhou Tang, Yaliang Li, and Kai Lei. 2018. CNN: Constructing a clinical Bayesain network based on data from the electronic medical record. Journal of biomedical informatics 88 (2018), 1–10.

[12] Ying Shen, Lixiu Zhang, Jin Zhang, Min Yang, Buzhou Tang, Yaliang Li, and Kai Lei. 2018. CNN: Constructing a clinical Bayesain network based on data from the electronic medical record. Journal of biomedical informatics 88 (2018), 1–10.

[13] Ying Shen, Lixiu Zhang, Jin Zhang, Min Yang, Buzhou Tang, Yaliang Li, and Kai Lei. 2018. CNN: Constructing a clinical Bayesain network based on data from the electronic medical record. Journal of biomedical informatics 88 (2018), 1–10.

[14] Ying Shen, Lixiu Zhang, Jin Zhang, Min Yang, Buzhou Tang, Yaliang Li, and Kai Lei. 2018. CNN: Constructing a clinical Bayesain network based on data from the electronic medical record. Journal of biomedical informatics 88 (2018), 1–10.

[15] Ying Shen, Lixiu Zhang, Jin Zhang, Min Yang, Buzhou Tang, Yaliang Li, and Kai Lei. 2018. CNN: Constructing a clinical Bayesain network based on data from the electronic medical record. Journal of biomedical informatics 88 (2018), 1–10.

[16] Ying Shen, Lixiu Zhang, Jin Zhang, Min Yang, Buzhou Tang, Yaliang Li, and Kai Lei. 2018. CNN: Constructing a clinical Bayesain network based on data from the electronic medical record. Journal of biomedical informatics 88 (2018), 1–10.

[17] Ying Shen, Lixiu Zhang, Jin Zhang, Min Yang, Buzhou Tang, Yaliang Li, and Kai Lei. 2018. CNN: Constructing a clinical Bayesain network based on data from the electronic medical record. Journal of biomedical informatics 88 (2018), 1–10.

[18] Ying Shen, Lixiu Zhang, Jin Zhang, Min Yang, Buzhou Tang, Yaliang Li, and Kai Lei. 2018. CNN: Constructing a clinical Bayesain network based on data from the electronic medical record. Journal of biomedical informatics 88 (2018), 1–10.