医療文書の見落とし軽減に向けた重要所見判定手法の妥当性評価

内堀優也[1] 亀谷聡[1]
田中良一[2] 鈴木智大[2] 田村明生[2]
[1]株式会社インテック [2]岩手医科大学
{uchibori_yuuya,kamegai_satoshi}@intec.co.jp
{rtanaka,tsuzuki,akahane}@iwate-med.ac.jp

概要

医師による医療文書の重要所見の見落としを防ぐ仕組みづくりが重要視されている．従来の固有表現抽出による所見判定手法では，検出された所見のうち重要とされる所見の割合が低く，検出された所見すべてを医師に通知するのでは，医師の負担軽減には不十分であった．そこで我々が実施した先行研究[1]では,固有表現抽出と文書分類を組み合わせて重要所見を判定する手法を考案した．本稿では，考案手法を複数の読影医(3 名)により評価し，考案手法の有効性を確認した．岩手医科大学で作成した読影レポートを対象に評価を実施した結果，マイクロ平均 f1 スコア 0.862 を達成し，重要所見を判定する手法としての有効性を確認した．また，ルールベース判定を導入することで，汎化性能を維持しつつ重要所見の確実な判定が可能となった．

1 はじめに

1.1 研究背景

病院等の医療現場においては，患者を担当する医師（依頼医）と検査画像の読影を行う医師（読影医）が異なることが大病院や大学病院では一般的である．依頼医は患者の診療に必要な情報を得るために読影医に CT などの検査画像の読影を依頼し,読影医が検査画像の読み解き及び診断を行い，所見の有無を記述した読影レポートを作成する．読影レポートに記述される内容の例を図 1 に示す[2].依頼医はこの読影レポートを基に，所見の有無を確認し，患者の診療に活かす．

右肺下葉S6 に境界不明瞭な約12mm のすりガラス状結節を認めます。1 年前の画像と比較して、約9mm → 約12mm と明らかに増大しています。微小浸潤部分は明瞭でありませんが、微小浸潤性腺癌（MIA）を疑います。

図 1 読影レポートの内容例

しかし，このプロセスにおいて，依頼医が重要所見を見逃す事例が発生しているため，読影レポートに対する確認不足や見落としを防ぐための仕組みづくりが重要視されている[3].

こうした背景を受けて，読影レポートに記載されている重要所見を依頼医へ通知する仕組みが考案されている．読影医が読影レポート作成時に重要所見を判定する手法が提案されているが，読影医の協力を得られない場合や読影医の負担の大きさが課題となる[4]．一方，用語辞書を作成して，ルールベースで重要所見を自動判定する手法が提案されているが,辞書にない用語の判定や曖昧な表現への対応，用語辞書の管理の負担が課題となる[5].

近年,固有表現抽出モデルに BERT や ChatGPT などの大規模言語モデルを用いて，読影レポートの所見を判定する手法が提案されており，所見を高い性能で判定できる可能性がある[6,7]．しかしながら，我々が行った調査では，読影レポート中に認められた所見の中で，重要といえる所見は全体の 30%以下であった．そのため，これら手法で判定した所見を医師に通知することは，重要ではない所見が多く通知されるため，医師の重要所見の見落としを防ぐ仕組みとしては不十分である．

1.2 我々が実施した先行研究

我々が実施した先行研究[1]では,重要所見を判定するために，固有表現抽出モデルで所見の有無を判定し，文書分類モデルで所見の重要度を判定する手法を提案した．岩手医科大学で作成した読影レポートで評価した結果，1 つの固有表現抽出モデルで判定するよりも高い精度で重要所見を判定できることがわかっている．

1.3 類似する先行研究との違い

本研究と類似する先行研究[8,9,10]では，読影レポートを対象にモデルを用いて所見検出や重要度判

定が行われている．これらの研究ではモデルが着目した文字列の傾向は示されるものの，重要所見の直接的な抽出には至っていない．

本研究では，固有表現抽出と文書分類を分けて行う手法を提案しており，所見と重要度を明確に抽出できる点で先行研究[8, 9, 10]と差別化される．これにより，依頼医へ重要所見をよりわかりやすく，正確に通知可能であり，この点において新たな意義を持つ．

1.4 本研究の目的と貢献

本研究の目的は，我々が実施した先行研究[1]で提案した固有表現抽出と文書分類を組み合わせた手法の有効性を複数の読影医により評価することである．

評価を実施する過程で，固有表現抽出結果を活用したルールベース判定を文書分類部分に適用する改良を加えた．

本研究の主な貢献は以下の通りである．

1. 複数の読影医（3 名）による評価を実施し，提案手法の有効性を実証した点．
2. 固有表現抽出の結果を活用したルールベース判定を文書分類部分に導入することで，汎化性能を落とさずに捉えたい重要所見を確実に判定可能にした点．

2 提案手法

読影レポートから重要所見のみを精度よく判定するため．我々が実施した先行研究[1]では所見有無判定と重要度判定を 2 段階に分ける手法を提案した．本研究では、この手法における重要度判定部分において、所見有無判定の結果を活用したルールベースの判定手法を導入した（図 2）．

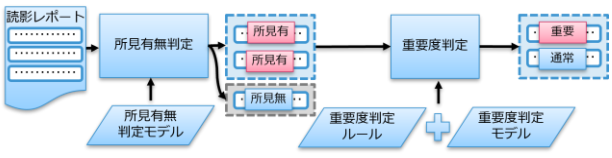

図 2 提案手法の概要

2.1 固有表現抽出（所見有無判定）

所見有無判定では，読影レポート中の所見の状態を判定する．本研究における所見とは，病名や症状，医療器具などを指す．これら所見について表 1 に示す 3 種類のいずれの状態か，固有表現抽出モデルで判定する．

表 1 所見有無の定義

状態	定義	記載例
認められる	ある病名などの存在が検査画像の中で実際に認められる状態。	○○認めます ○○散在
疑いがある	ある病名などの存在が検査画像の中で疑われる状態、鑑別疾患として提案されている場合。	○○否定はできません ○○は鑑別にあがる
認められない	ある病名などの存在が検査画像の中で否定された場合。	○○は消失 ○○術後

3 種類の状態の定義やアノテーションは，Real-MedNLP のアノテーションガイドライン[5]を参考に実施する．固有表現抽出モデルには，固有表現抽出タスクにおいて高い精度が報告されている Bi-LSTM+CRF[11, 12]を用いる．

2.2 文書分類（重要度判定手法）

重要度判定では，所見を含む文の重要度を文書分類モデルで判定する．本研究における重要度とは，臨床的に患者にとって不利益があるかどうかを表し，高いほど不利益があり重要とみなす（表 2）．アノテーションは，読影医が本定義を基に実施した．文書分類モデルには，文書分類タスクにおいて高い精度が報告されている BERT[13]を用いる．

表 2 重要度の定義

重要度	定義	記載例
5	確実に重要な所見を含む。	腎腫瘍が疑われます。
4	重要な可能性がある所見を含む。	胸水が悪化しています。
3～1	重要とはいえない所見を含む。	少量胸水あり。
0	正常な所見を含む。	門脈は開存しています。

2.3 ルールベース判定の活用

読影医による重要度判定結果の評価において，許容できない誤判定が存在した．そこで文書分類部分に，確実に判定したい重要所見を判定可能にするため，固有表現抽出結果を活用したルールベース判定を適用している．ルール適用手順は以下の通りである．

1. 所見有無判定で認められるまたは疑いがあると判定された文字列を抽出．
2. 上記文字列にキーワードが含まれる場合，所定の重要度を当該文に付与する．この場合，重要度判定モデルの結果は考慮しない．

本ルールを適用する際のキーワードと重要度の例を表 3 に示す．

表 3 キーワードと重要度の例

タイプ	キーワード	重要度
症状	MS病変、腎腫瘍、有棘細胞癌	5
	原発性肺癌、多房性嚢胞性腫瘤、食道壁肥厚	4
症状＋状態	主膵管拡張＋増大、多房性嚢胞＋増大	4

3 評価実験

提案手法を読影レポートに適用した際の性能を評価する．

3.1 データセット

岩手医科大学で過去に作成された読影レポート 350 件の所見欄を抽出してデータセットとして利用した．表 4 にデータセットの統計情報を示す．

表 4 データセットの統計情報

読影レポート件数	350
1件あたりの文字数（平均）	185.08
1件あたりの文数（平均）	8.49

データセットに対して，所見有無，および重要度のアノテーションを実施した．アノテーション数をそれぞれ，表 5，表 6 に示す．

表 5 所見有無アノテーション数

アノテーション数	3,212
認められる	1,597
疑いがある	421
認められない	1,194

表 6 重要度アノテーション数

アノテーション数	1,709
5	147
4	343
3～1	994
0	225

3.2 評価方法

固有表現抽出モデルの判定結果の正否判定は，2 通りの方法（完全一致，部分一致）を用いる．「完全一致」は，モデル判定とアノテーションの所見有無および位置が完全一致した場合のみ正解と見なす．一方，「部分一致」は，所見有無が一致し，位置が部分一致する場合も正解と見なす．両者の違いを表 7 に示す．評価指標には F1 を用いて，10-fold 交差検証法で作成した各モデルの評価値の平均値を算出して評価する．

表 7 正誤判定の違い

正否判定方法		アノテーション		モデル判定		文
完全一致	部分一致	所見有無	位置	所見有無	位置	
正解	正解	認められる	脂肪肝	認められる	脂肪肝	脂肪肝の状態です。
不正解	正解	疑いがある	肺転移	疑いがある	転移	肺転移を疑います。

文書分類モデルの判定結果の正否判定は，2 通りの方法（個別，集約）を用いる．「個別」は，重要度ごとに正否判定を行う．一方，「集約」は，表 8 に示すように重要度を集約して，正否判定を行う．

表 8 集約方法

重要度（個別）	集約
5	4以上
4	
3～1	3以下
0	

集約する理由は，重要所見を含む文と含まない文に対する，それぞれの重要度判定の性能を評価し易くするためである．両者の違いを表 9 に示す．評価指標には F1 を用いて，10-fold 交差検証法で作成した各モデルの評価値の平均値を算出する．

表 9 正否判定方法の違い

正否判定方法		アノテーション（重要度）	モデル判定（重要度）	文
個別	集約			
正解	正解	3～1	3～1	脂肪肝の状態です。
不正解	正解	5	4	肺転移を疑います。

所見有無判定結果と重要度判定結果を組み合わせた場合の評価を行う．まず，所見有無判定を用いて「認められる」または「疑いがある」所見を含む文を抽出し，抽出した文に対して重要度判定を行う（図 3）．評価指標は，文書分類モデルの判定結果と同様に F1 を用いる．

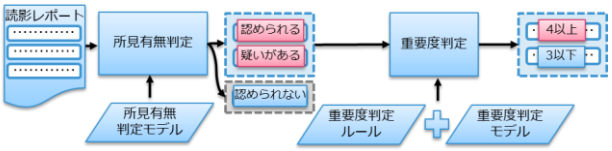

図 3 提案手法の概要

4 評価結果

4.1 所見有無判定手法結果

「完全一致」，「部分一致」で正否判定した場合の所見有無判定結果を表 10 に示す．マイクロ平均 F1 値は「部分一致」の方が「完全一致」よりも高く，

「認められない」と「認められる」に比べて，「疑いがある」の F1 値が低い．

表 10 所見有無判定結果

	完全一致			部分一致		
	Precision	*Recall*	*F1*	*Precision*	*Recall*	*F1*
マイクロ平均	0.906	0.923	0.914	0.974	0.980	0.977
認められる	0.885	0.911	0.898	0.972	0.987	0.979
疑いがある	0.845	0.873	0.858	0.954	0.945	0.949
認められない	0.957	0.957	0.957	0.986	0.982	0.984

部分的にでも所見有無を判定できれば，依頼医へ通知できる．また，所見の詳細は依頼医が目視により把握できるため，本研究では，「部分一致」の精度が重要となる．「部分一致」の方法で算出した，所見有無判定性能としては，マイクロ平均 F1 値 0.977 を示した．

4.2 重要度判定手法結果

「個別」，「集約」で正否判定した場合の重要度判定結果を表 11 に示す．「重要度 5」および「重要度 4」よりも，「重要度 4 以上」に集約した方が F1 値は高い．重要所見の見落としを防ぐという観点では，重要度 4 以上を正しく 4 以上と判定することが重要となる．「重要度 4 以上」の方法で算出した，重要度判定性能としては，F1 値 0.766 を示した．

表 11 重要度判定結果

	個別			集約		
	Precision	*Recall*	*F1*	*Precision*	*Recall*	*F1*
マイクロ平均	0.774	0.774	0.774	0.866	0.866	0.866
重要度5	0.566	0.685	0.620			
重要度4	0.629	0.512	0.565			
重要度4以上				0.798	0.737	0.766
重要度3～1	0.834	0.881	0.857			
重要度3以下				0.892	0.921	0.907
重要度0	0.854	0.788	0.820			

ルール適用前の「個別」，「集約」で正否判定した場合の重要度判定結果を表 12 に示す．ルール適応前後での性能差はわずかであり，読影医が許容できないとした誤判定を性能を落とすことなく判定する可能性を示した．

表 12 重要度判定結果（ルール適用前）

	個別			集約		
	Precision	*Recall*	*F1*	*Precision*	*Recall*	*F1*
マイクロ平均	0.772	0.772	0.772	0.864	0.864	0.864
重要度5	0.564	0.678	0.616			
重要度4	0.625	0.504	0.558			
重要度4以上				0.796	0.730	0.761
重要度3～1	0.831	0.881	0.855			
重要度3以下				0.890	0.921	0.905
重要度0	0.854	0.788	0.820			

また，ルール適用により正しく判定可能になった例を表 13 に示す．所見有無判定にて疑いがあると判定された所見を含む文は，同様の所見を認められると判定された文よりも重要度判定結果が低くなる傾向があることが示唆される．

表 13 ルール適用例

対象例	モデル識別結果	ルール適用後
こちらも**原発性肺癌**の可能性が疑われます。	3~1	4
子宮体部右側壁に接して長径42mm程度の**多房性嚢胞性腫瘤**を認めます(図1)。	3～1	4
胸部上部食道壁肥厚しています。	3~1	4

4.3 両判定結果組み合わせ結果

所見有無判定と重要度判定を組み合わせた場合の評価結果を表 14 に示す．

マイクロ平均 F1 値 0.862 であり，実用的な性能であることを示した．

表 14 両判定結果組み合わせ結果

	両判定組み合わせ		
	Precision	*Recall*	*F1*
マイクロ平均	0.862	0.862	0.862
重要度4以上	0.788	0.743	0.765
重要度3以下	0.892	0.914	0.903

5 おわりに

本稿では，読影レポート中の重要所見を判定するために，我々が実施した先行研究[1]で提案した固有表現抽出と文書分類を組み合わせた手法の有効性を複数の読影医により評価した．岩手医科大学で作成した読影レポートで評価した結果，高い精度で重要所見を判定できることがわかった．また，ルールベース判定を導入することで，汎化性能を維持しつつ重要所見の確実な判定が可能となることも確認した．今後は，本手法の判定結果を依頼医に通知することの有効性を実証し，医師の重要所見の見落とし防止への活用に向けた取り組みを進める．依頼医に通知する際の画面イメージ例を図 4 に示す．

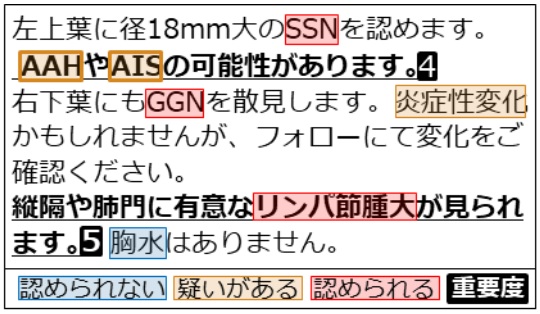

図 4 見落とし防止への活用案

利益相反

本研究は岩手医科大学と株式会社インテックの共同研究として実施された．本研究に関連して，岩手医科大学と株式会社インテックは共同で特許を出願中である．

参考文献

[1] 亀谷聡，北橋竜雄，田中良一：医療文書の見落とし軽減に向けた重要所見判定手法の提案，人工知能学会，人工知能学会全国大会 39 回，2025.

[2] MedNLP. 公開読影レポートデータセット_202010, 2020. https://sociocom.naist.jp/download/reading-reports_202010. Accessed: 2025-01-21.

[3] 厚生労働省．医療安全に資する病院情報システムの機能を普及させるための施策に関する研究，2018 . https://mhlw-grants.niph.go.jp/project/27426. Accessed: 2025-01-21.

[4] 藤井歩美，武田理宏，村田泰三，橋本麻紀子，向井 頼貴，真鍋史朗，松村泰志：画像診断レポートの見落とし防止に向けた対策と有効性の検証，医療情報学連合大会論文集 39 回，2019.

[5] 杉本賢人，孫逸樵，和田聖哉，小西正三，岡田佳築，松村 泰志，武田理宏：自然言語処理を用いた読影レポートからの重要所見の自動抽出に関する取り組み，医療情報学連合大会論文集 43 回，2023.

[6] 矢田 竣太郎，田中 リベカ，Fei Cheng，荒牧英治，黒橋 禎夫:汎用的な臨床医学テキストアノテーション仕様およびガイドラインの策定：重篤肺疾患ドメインに着目して，自然言語処理， 2022.

[7] 西山智弘，柴田大作，宇野裕，辻川剛範，北出祐，久保雅洋，矢田竣太郎，若宮翔子，荒牧英治：生成モデルは医療文書の固有表現抽出に使えるか?，言語処理学会年次大会発表論文集 30 回，2024.

[8] Kento SUGIMOTO, Shoya WADA, Shozo KONISHI, Katsuki OKADA, Shirou MANABE, Yasushi MATSUMURA, Toshihiro TAKEDA: Classification of Diagnostic Certainty in Radiology Reports with Deep Learning, Stud Health Technol Inform. 2024.

[9] Junya Sato, Kento Sugimoto, Yuki Suzuki, Tomohiro Wataya, Kosuke Kita, Daiki Nishigaki, Miyuki Tomiyama, Yu Hiraoka, Masatoshi Hori, Toshihiro Takeda, Shoji Kido, Noriyuki Tomiyama: Annotation-free multi-organ anomaly detection in abdominal CT using free-text radiology reports: a multi-centre retrospective study, EBioMedicine. 2024.

[10] Tomohiro Wataya, Azusa Miura, Takahisa Sakisuka, Masahiro Fujiwara, Hisashi Tanaka, Yu Hiraoka, Junya Sato, Miyuki Tomiyama, Daiki Nishigaki, Kosuke Kita, Yuki Suzuki, Shoji Kido, Noriyuki Tomiyama: Comparison of natural language processing algorithms in assessing the importance of head computed tomography reports written in Japanese, Jpn J Radiol. 2024.

[11] Zhiheng Huang, Wei Xu, Kai Yu: Bidirectional LSTM-CRF Models for Sequence Tagging, 2015.

[12] Ken Yano: Neural Disease Named Entity Extraction with Character-based BiLSTM+CRF in Japanese Medical Text, 2018.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2019.