Radiography 28 (2022) 674–683

Contents lists available at ScienceDirect

Radiography

journal homepage: www.elsevier.com/locate/radi

Development of lumbar spine MRI referrals vetting models using machine learning and deep learning algorithms: Comparison models vs healthcare professionals

A.H. Alanazi [a, b, *], A. Cradock [a], L. Rainford [a]

[a] Radiography and Diagnostic Imaging, School of Medicine, University College Dublin, Ireland
[b] Society of Artificial Intelligence in Healthcare, Riyadh, Saudi Arabia

ARTICLE INFO

Article history:
Received 24 March 2022
Received in revised form
28 April 2022
Accepted 24 May 2022
Available online 11 June 2022

Keywords:
Natural language processing
Machine learning
Deep learning
Referrals' appropriateness
Magnetic resonance imaging

ABSTRACT

Introduction: Referrals vetting is a necessary daily task to ensure the appropriateness of radiology referrals. Vetting requires extensive clinical knowledge and may challenge those responsible. This study aims to develop AI models to automate the vetting process and to compare their performance with healthcare professionals.

Methods: 1020 lumbar spine MRI referrals were collected retrospectively from two Irish hospitals. Three expert MRI radiographers classified the referrals into indicated or not indicated for scanning based on iRefer guidelines. The reference label for each referral was assigned based on the majority voting. The corpus was divided into two datasets, one for the models' development with 920 referrals, and one included 100 referrals used as a held-out for the final comparison of the AI models versus national and international MRI radiographers. Three traditional models were developed: SVM, LR, RF, and two deep neural models, including CNN and Bi-LSTM. For the traditional models, four vectorisation techniques applied: BoW, bigrams, trigrams, and TF-IDF. A textual data augmentation technique was applied to investigate the influence of data augmentation on the models' performances.

Results: RF with BoW achieved the highest AUC reaching 0.99. CNN model outperformed Bi-LSTM with AUC = 0.98. With the augmented dataset, the performance significantly improved with an increase in F1 scores ranging from 1% to 7%. All models outperformed the national and international radiographers when compared on the hold-out dataset.

Conclusion: The models assigned the referrals' appropriateness with higher accuracies than the national and international radiographers. Applying data augmentation significantly improved the models' performances.

Implications for practice: The outcomes suggest that the use of AI for checking referrals' eligibility could serve as a supporting tool to improve the referrals' management in radiology departments.

Introduction

Several studies have investigated the appropriateness of Lumbar Spine Magnetic Resonance Imaging (LSMRI) referrals through analysis of their compliance to the international guidelines, such as the American College of Radiology Appropriateness (ACR)[1] and Royal College of Radiologists (RCR).[2] The range of unjustified LSMRI scans has been identified to be between 10% and 60%.[3–6] Common reasons of ordering inappropriate examinations including lack of awareness of the applied guidelines and seeking patient satisfaction.[7] The rise in unjustified cases has prompted health organisations to seek effective solutions to combat this phenomenon.[8]

Two approaches are proposed to limit unjustified referrals. One is to offer educational courses related to the referrals' appropriateness for radiology employees, however this may have a financial impact on radiology departments' budgets.[9] The second approach is to use clinical decision-support tools, which provides referring physicians with real-time feedback on the referral's appropriateness. This approach has been shown to reduce unjustified examinations,[10–12] but requires training on how to use such a tool, which employ questions and checklists before manual order entry. To overcome these constraints, automating the process of referrals

* Corresponding author. South Central Building Apartment 29, Dublin, Sandyford, D18 RW02, Ireland.
E-mail address: Ali.alanazi@ucdconnect.ie (A.H. Alanazi).

https://doi.org/10.1016/j.radi.2022.05.005

Abbreviations:

ACR	American College of Radiology
AI	Artificial Intelligence
AUC	Area Under the Curve
Bi-LSTM	Bidirectional Long Short-Term Memory
BoW	Bag of Words
CNN	Convolutional Neural Network
DA	Data Augmentation
DL	Deep Learning
LR	Logistic Regression
LSMRI	Lumbar Spine Magnetic Resonance Imaging
ML	Machine Learning
MRI	Magnetic Resonance Imaging
NLP	Nature Language Processing
NLTK	Natural Language Toolkit
OOV	Out of Vocabulary
RBF	Radial Basis Function
RCR	Royal College of Radiologists
RF	Random Forest
SPSS	Statistical Package for the Social Sciences
SVM	Support Vector Machine
TF-IDF	Term Frequency-Inverse Document Frequency

vetting by using machine learning based Natural Language Processing (NLP) may offer an efficient and cost-effective approach.

Advances in machine learning have demonstrated significant promise in auto-categorisation of radiology texts. For instance, Trivedi et al.,[13] developed a decision-making model for auto-assigning the need for contrast for musculoskeletal MRI examinations based on the clinical indications written in the requests with an accuracy of 83%. In another study,[14] the authors built various models for auto-classification of MRI knee reports into normal and abnormal findings. They experimented with SVM and reported an F1 score of 0.90. Moreover, in Singapore,[15] different models were built for auto-auditing the appropriateness of brain MRI requests and an F1 score of 0.94 was reported for the XGBoost algorithm.

In this study, we experimented with Support Vector machine (SVM), Logistic Regression (LR), Random Forest (RF), Convolutional Neural Network (CNN), and Bi-directional Long Short-Term Memory (Bi-LSTM). We integrated our deep neural models (CNN and Bi-LSTM) with pre-trained embedding model developed by Facebook called Fasttext (2 million word vectors trained on common crawl datasets: https://fasttext.cc/docs/en/english-vectors.html). A pre-developed text augmentation technique was implemented to adjust data imbalance and investigate its influence on the models' performances. Furthermore, we analysed the performance of four vectorisation techniques implemented on the original and the augmented datasets with the traditional models: BoW, bigram, trigram, and TF-IDF. Finally, an unseen dataset was used to compare the best models versus eight national and international MRI radiographers.

Methods and materials

Study design

The study aimed to develop AI models to auto-classify LSMRI referrals into indicated and not indicated for scanning and to be directly compared versus professionals. Five algorithms were used: SVM, LR, RF, CNN, and Bi-LSTM. Two datasets employed; the original dataset including referrals gathered retrospectively from two hospitals, private and public, and an augmented version of the original dataset. Several vectorisation techniques were experimented with to explore the highest performing technique. Reporting of this paper follows the Checklist for Artificial Intelligence in Medical Imaging (CLAIM).[16]

Data

Data corpus

Institutional ethical approvals were confirmed and related agreements and data gathering permissions were obtained. Clinical indications were extracted in the form of "referral texts" from 1020 LSMRI referrals from radiology departments in two Irish hospitals, public and private, 593 and 427, respectively. The data was collected from two hospitals to incorporate a variety of language styles in the corpus. To ensure referrals were clear and anonymised, all irrelevant material, colloquialisms and material relating to the local service and patients' identities were edited.

Creation of the reference labels

Three MRI radiographers with experience ranging from 17 to 20 years were requested to assign the appropriateness of the 1020 LSMRI referrals as indicated or not indicated for LSMRI independently and based on iRefer guidelines.[3] iRefer was the choice for this study as the participating radiographers worked within an environment that routinely used iRefer. There are substantial similarities between iRefer, ACR guidelines, iESR and the Australian/New Zealand guidelines with respect to lumbar spine referral guidance. After obtaining all categorisations from the three radiographers, the reference labels were created based on the majority vote for each referral. The reference included 119 (11.6%) not indicated, and 901 (88.4%) indicated referrals.

Division of the corpus

After identification of the referrals' labels in the *creation of the reference labels* section, the corpus was split into two datasets. The first dataset, called the original dataset, included 920 referrals (820 indicated and 100 not indicated) and was utilised for models' development. The second dataset included 100 referrals (81 indicated and 19 not indicated) and was utilised as a held-out dataset for the final comparison with radiographers.

Creation of the augmented dataset

To increase the number of the not indicated referrals in the original dataset, a pre-developed swapping words model (available online: https://www.kaggle.com/shonenkov/nlp-albumentations/) was applied to only the not indicated referrals in the original dataset. For multi-augmentations of a single text, we applied the process 3 consecutive times, and in each time, a new text was produced. For each new text production, the model requires a manual entry of the variables *swap_distance* and *swap_probability*, which are responsible for distance and probability of swapping one word. Fig. 1 shows an example of a not indicated request that was augmented three times. Consequently, the number of not indicated referrals in the augmented dataset increased from 100 to 400, making the total referrals for this dataset 1220 (820 indicated and 400 not indicated).

Data pre-processing

Text cleaning

The texts were cleaned using NLTK toolkit. Uninformative words (Stop-words), such as *in, on,* and *the,* were removed using the *stopwords* removal function. A customised punctuations removal

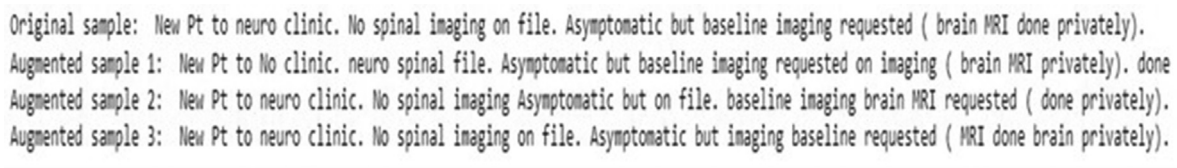


Figure 1. An example of not indicated LSMRI referral before and after the augmentation applied using swapping words method.

function was built to remove punctuations, and we preserved punctuations with temporal connotations, such as "<, >,/".

Texts vectorisation

Traditional machine Learning models can not comprehend written words and each text must be fed to the model in a form of vector containing numerical values. In this work, the original and augmented datasets were experimented with three bag of words techniques: bag-of-words BoW, bigram, trigram, and one weighting technique named Term Frequency-Inverse Document Frequency TF-IDF. Hence, four feature vectorisation approaches were implemented.

Development of AI models

The models were developed using an open-source library, Google-Colab. A split ratio of 70–30 was used to train and test the models. Fig. 2 shows the pipeline of the project.

Traditional machine learning models

SVM, LR, and RF were chosen for models' development. Each algorithm was experimented with the original and augmented datasets. In addition, each algorithm was experimented with the four vectorisation techniques mentioned in the *Texts vectorisation* section. Hence, 24 traditional models were developed. In each model, the grid search was applied to identify the optimal parameters' values. Three kernels experimented with SVM models: linear, polynomial, and the radial basis function kernel (RBF). Other SVM parameters, including C regularization and gamma, were tested with various values. For LR models, C regularization and L1, L2 penalties were searched. Four parameters were searched in the RF models: max_depth, max_feature, min_sample_leaf, and n_estimator

Deep neural network models

CNN and Bi-LSTM algorithms were selected for the models' development. Each algorithm was experimented with both datasets. Hence, 4 deep neural models were developed.

CNN models included the Fasttext word embedding which was used to convert words into vectors with 300 dimensions, followed by a single CNN layer with 50 neurons and 3×3 convolutions and Relu activation. After CNN layer, global max-pooling was used to take the maximum value of the convolutions. A dense layer with 50 neurons and Relu activation was used to flatten the output of the max-pooling. The final layer, the output layer, included one neuron with sigmoid activation to output binary values. Both models were compiled using Adam optimizer with learning rate = 2e-2, and binary cross entropy loss function, and manually optimised using batch size = 180, epochs = 30, and 0.5 dropout after each layer for the model trained on the original dataset and 0.3 dropout after each layer for the model trained on the augmented dataset.

For Bi-LSTM models, Fasttext word embedding was used, followed by a single Bi-LSTM layer with 50 neurons for the model trained on the original dataset and 10 neurons for the model trained on the augmented dataset. After Bi-LSTM layer, a dense layer with 50 neurons and Relu activation was used for the model trained on the original dataset and 80 neurons with Relu activation for the model trained on the augmented dataset. The output layer contained one neuron with sigmoid activation. Both models were compiled using Adam optimizer with learning rate = 2e-2, and binary cross entropy loss function, and manually optimised using batch size = 180, epochs = 30, and 0.5 dropout after each layer.

Performance evaluation and selection of the highest performing models

To assess each model's performance, accuracy, precision, and recall were calculated. To select the highest models, the receiver operating characteristic curves ROC were plotted, and the area under the curve (AUC) was used with F1 score.

Final comparison with national and international radiographers

Eight (n = 8) MRI radiographers with experience ranging from 7 to 35 years were recruited to categorise referrals in the held-out dataset: two non-reporting radiographers from Ireland, two non-reporting from Malta, two non-reporting from the UK, and two reporting from the UK. All radiographers are currently/had been using iRefer. The radiographers were asked to assign the referrals' appropriateness as indicated or not indicated for MRI independently based on iRefer guidelines. The highest performing models were re-tested using the held-out dataset. Both radiographers and models' outputs were benchmarked to the reference labels of the held-out referrals (Fig. 2). Furthermore, Cohen's kappa analysis was performed to measure the agreement between each categorisation and the reference labels. Agreement levels follows levels suggested by McHugh (2012),[17] in which kappa <0.00 is interpreted as no agreement, 0.01–0.20 slight, 0.21–0.40 fair, 0.41–0.60 moderate, 0.60–0.80 substantial, and 0.81–1.00 as perfect agreement.

Results

Traditional machine learning models

The LR model with TF-IDF achieved the highest AUC and F1 score among models trained and tested on the original dataset, 0.761 and 0.942, respectively. LR was optimised with C regulariser = 27.82 and L2 penalty. With the augmented dataset, RF with BoW techniques (RF-BoW-DA, RF-Bigram-DA, RF-Trigram-DA) achieved the highest AUCs and F1 scores, reaching 0.996 and 0.993, respectively. RF with bigrams and trigrams were optimised with 400 estimators, 50 maximum depth, squared maximum features, and 1 minimum samples leaf. RF with BoW was optimised with 400 estimators, 70 maximum depth, squared maximum features, and 1 minimum samples leaf. Table 1 presents the results of all traditional models.

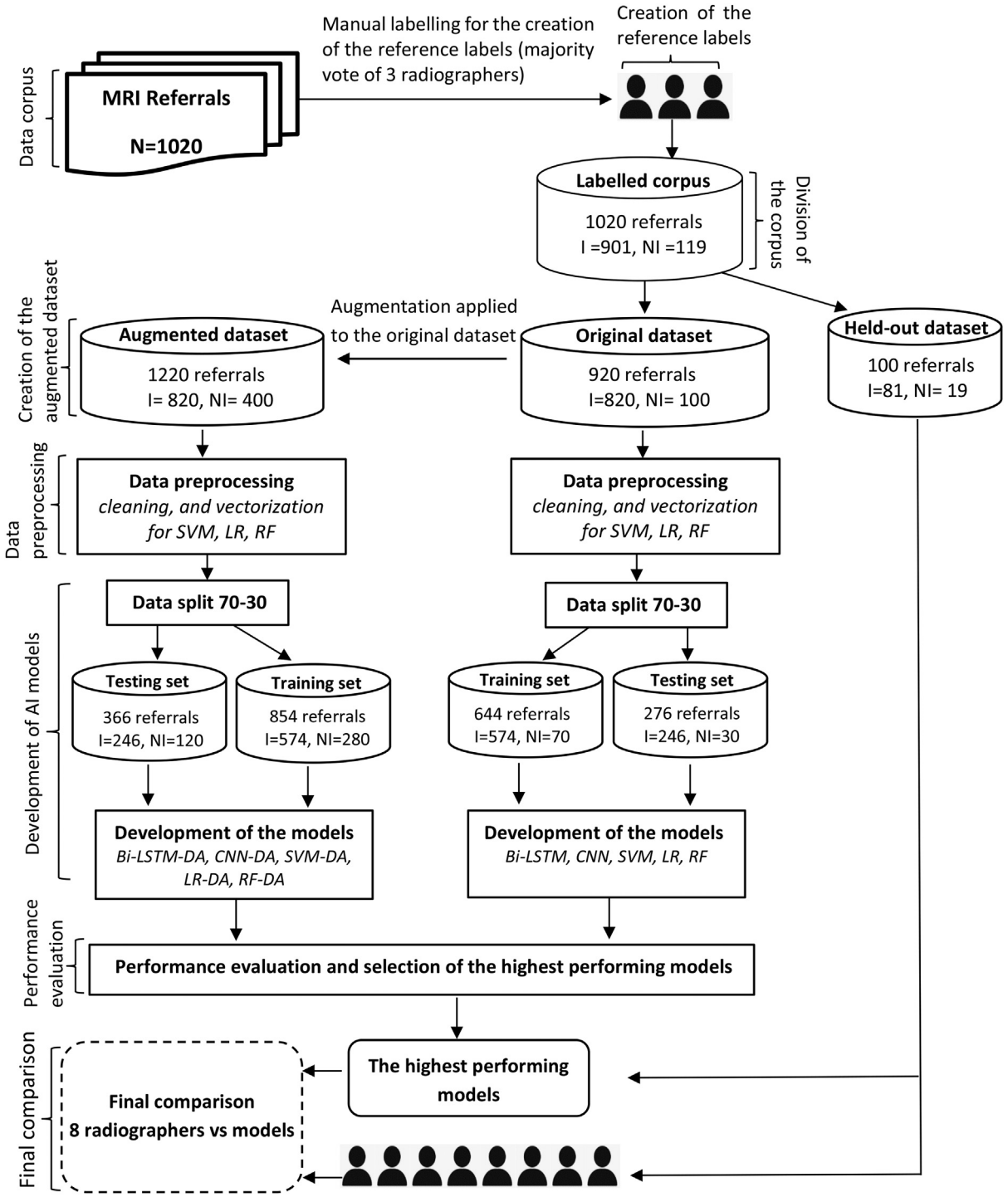


Figure 2. A diagram showing the pipeline for data pre-processing, models' development, and final comparison. Note: NI = not indicated, I = indicated for scanning.

Deep neural models

Both CNN and Bi-LSTM demonstrated similar performance with AUCs of 0.740 when trained and tested on the original dataset. With the augmented dataset, CNN-DA showed higher performance than Bi-LSTM-DA with AUC = 0.981 and F1 = 0.941. Table 1 presents the results of all deep neural models.

The highest performing models

The highest performing models are those trained and tested on the augmented dataset (RF–BoW–DA, RF–bigram–DA, RF–tri-gram–DA, CNN–DA, and Bi-LSTM–DA). Those models were selected based on their high AUCs and F1 scores. Figs. 3 and 4 show the AUC and the confusion matrix of each model.

Table 1
Training and testing results of all developed models using the original and augmented datasets.

Machine learning models	Training results					Testing results				
	Acc	Pre	Rec	F1	AUC	Acc	Pre	Rec	F1	AUC
SVM-BoW	0.94	0.93	1.00	0.967	0.999	0.89	0.89	1.00	0.942	0.711
SVM-Bigrams	0.94	0.93	1.00	0.967	0.999	0.89	0.89	1.00	0.942	0.711
SVM-Trigrams	0.94	0.93	1.00	0.967	0.999	0.89	0.89	1.00	0.942	0.711
SVM-TF-IDF	0.93	0.93	1.00	0.965	0.999	0.89	0.89	1.00	0.942	0.711
SVM-BoW-DA	0.99	1.00	0.98	0.993	0.999	0.97	0.98	0.97	0.979	0.971
SVM-Bigrams-DA	0.98	1.00	0.98	0.992	0.999	0.96	0.98	0.96	0.975	0.970
SVM-Trigrams-DA	0.98	1.00	0.98	0.992	0.999	0.96	0.98	0.96	0.975	0.970
SVM-TF-IDF-DA	0.99	0.99	0.99	0.995	0.999	0.92	0.97	0.92	0.945	0.968
LR-BoW	0.99	0.99	1.00	0.997	1.00	0.86	0.91	0.93	0.923	0.750
LR-Bigrams	0.99	0.99	1.00	0.997	1.00	0.86	0.90	0.94	0.926	0.687
LR-Trigrams	0.99	0.99	1.00	0.997	1.00	0.86	0.90	0.94	0.926	0.687
LR-TF-IDF	0.99	0.99	1.00	0.999	1.00	0.89	0.89	0.99	0.942	0.761
LR-BoW-DA	0.99	1.00	0.99	0.996	0.999	0.95	1.00	0.93	0.964	0.988
LR-Bigrams-DA	0.99	1.00	0.99	0.997	0.999	0.94	1.00	0.91	0.957	0.988
LR-Trigrams-DA	0.99	1.00	0.99	0.997	0.999	0.94	1.00	0.91	0.957	0.998
LR-TF-IDF-DA	0.99	0.99	0.99	0.996	0.999	0.97	0.98	0.97	0.981	0.997
RF-BoW	0.98	0.97	1.00	0.989	0.998	0.89	0.89	0.99	0.942	0.724
RF-Bigrams	0.98	0.97	1.00	0.989	0.999	0.88	0.89	0.99	0.940	0.756
RF-Trigrams	1.00	1.00	1.00	1.00	1.00	0.86	0.89	0.97	0.929	0.749
RF-TF-IDF	1.00	1.00	1.00	1.00	1.00	0.88	0.89	0.99	0.938	0.757
RF-BoW-DA	1.00	1.00	1.00	1.00	1.00	0.97	1.00	0.96	**0.983**	**0.995**
RF-Bigrams-DA	0.99	0.99	1.00	0.996	1.00	0.98	1.00	0.98	**0.991**	**0.996**
RF-Trigrams-DA	0.99	0.99	1.00	0.996	0.996	0.99	1.00	0.98	**0.993**	**0.996**
RF-TF-IDF-DA	0.99	0.99	1.00	0.999	1.00	0.97	0.97	0.98	0.979	0.988
Deep neural models										
CNN	0.95	0.95	0.99	0.974	0.970	0.86	0.88	0.98	0.929	0.739
CNN-DA	0.94	0.95	0.97	0.963	0.991	0.92	0.97	0.91	**0.941**	**0.981**
Bi-LSTM	0.98	0.98	0.99	0.990	0.999	0.85	0.89	0.95	0.921	0.740
Bi-LSTM-DA	0.93	0.95	0.95	0.953	0.983	0.91	0.96	0.91	**0.935**	**0.973**

Note: : Bright rows denote models trained and tested on the original dataset, shaded rows denote models trained and tested on the augmented dataset, Acc = accuracy, pre = precision, Rec = recall, F1 = F1 score, AUC = area under the receiver operating characteristic curve, SVM = support vector machine algorithm, LR = logistic regression algorithm, RF = random forest algorithm, CNN = convolutional neural network, Bi-LSTM = bi-directional neural network, BoW = bag-of word-feature extraction technique, Bigram = bigram feature extraction technique, Trigram = trigram feature extraction technique, TF-IDF = term frequency-inverse document frequency technique, DA = (Data Augmented: models trained on the augmented data).

Final comparison: models versus national and international MRI radiographers

The highest performing models from *the highest performing models* section performed higher than the eight radiographers when re-tested on the held-out dataset. Bi-LSTM-DA showed higher F1 than other models. Kappa showed higher agreement between the classifications of the models and the reference labels than between the radiographers and the reference labels (Table 2).

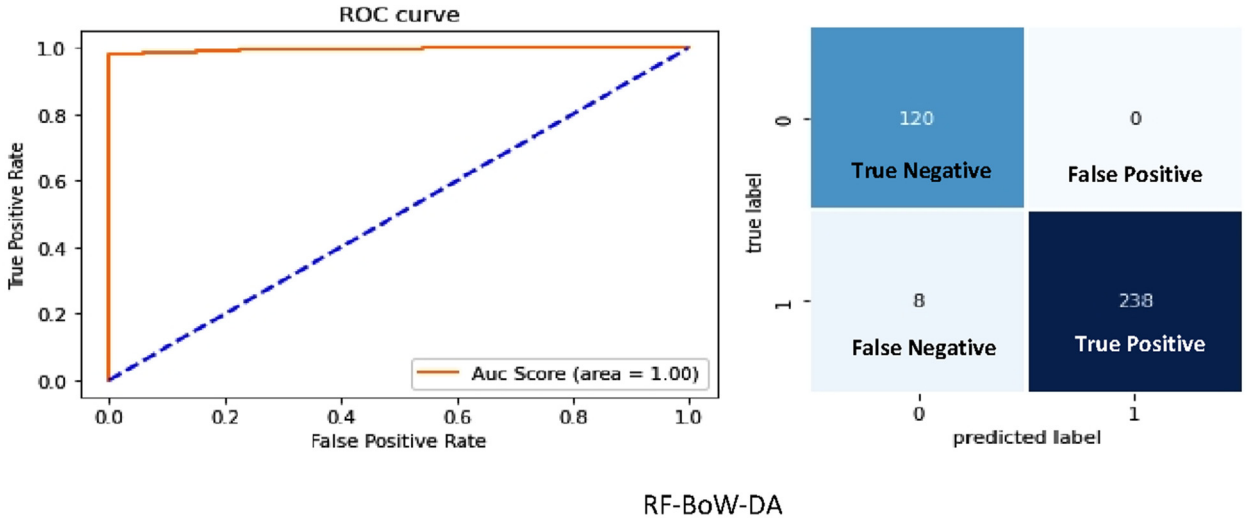
Discussion

We experimented with machine learning and deep neural models using several vectorisations and data augmentation techniques. The models demonstrated the ability for auto-vetting referrals' appropriateness. Outcomes suggest that using AI models for textual data in radiology referrals demonstrates potential support for radiology departments in referrals' management.

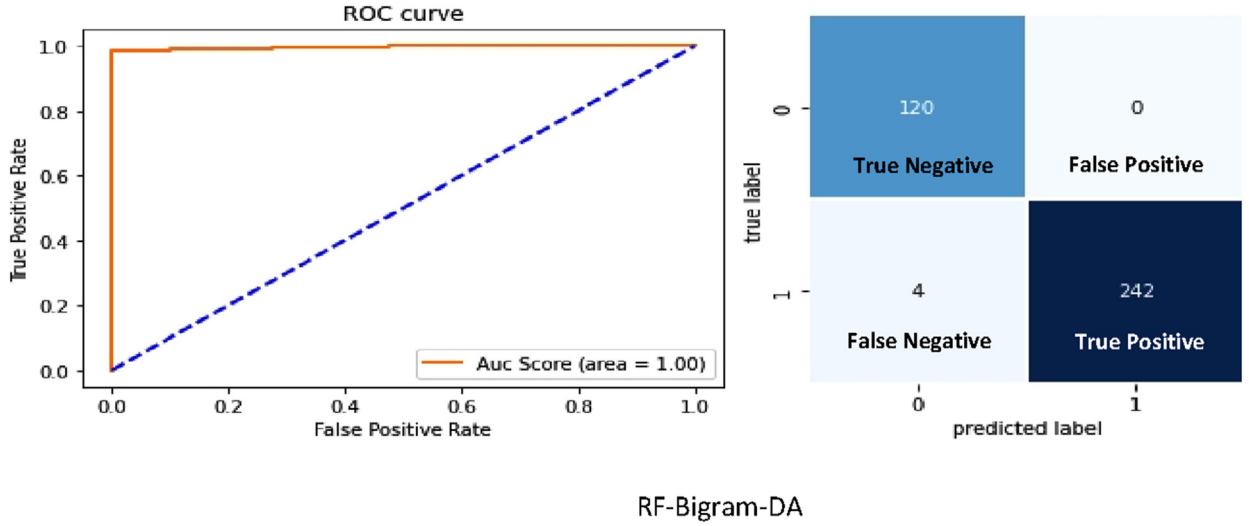
Numerous studies have aimed to apply machine learning ML and deep learning DL on radiology text to improve healthcare quality and to facilitate data retrieval.[18] Nevertheless, there is a promising future for the applications of ML and DL to improve healthcare management. Unjustified referrals continue to be problematic for radiology departments, and impact upon service endpoints, such as utilisation of resources and patient safety.

In the original dataset, there was severe class imbalance as there were 100 (10.8%) referrals labelled as not indicated and 820 (89.2%) labelled as indicated. The ratio of indicated over not indicated referrals was 8.2 (820/100 = 8.2). This imbalance is expected as referrals were randomly collected; in other words, the data imbalance is a direct result of the nature of the medical data.[19]

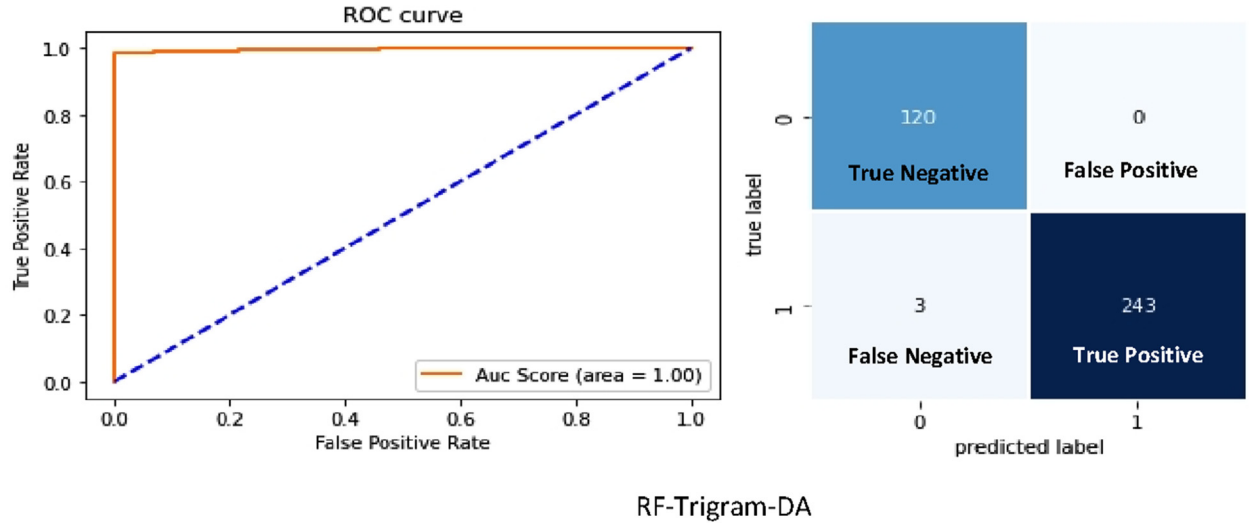
Imbalanced data for training a model can cause overfitting, in which the model performs substantially worse on the testing data than it did on the training data.[20] To better understand the effect imbalanced datasets, the models in this study experimented with the original dataset (imbalanced) and the augmented dataset, in which the ratio of indicated over the not indicated referrals was adjusted to be 2.05 (820/400 = 2.05). Data augmentation



RF-BoW-DA



RF-Bigram-DA



RF-Trigram-DA

Figure 3. AUC and confusion matrix of the highest performing traditional machine learning models with data augmentation. Note: There is ±0.05 change in the score of the plotted AUC.

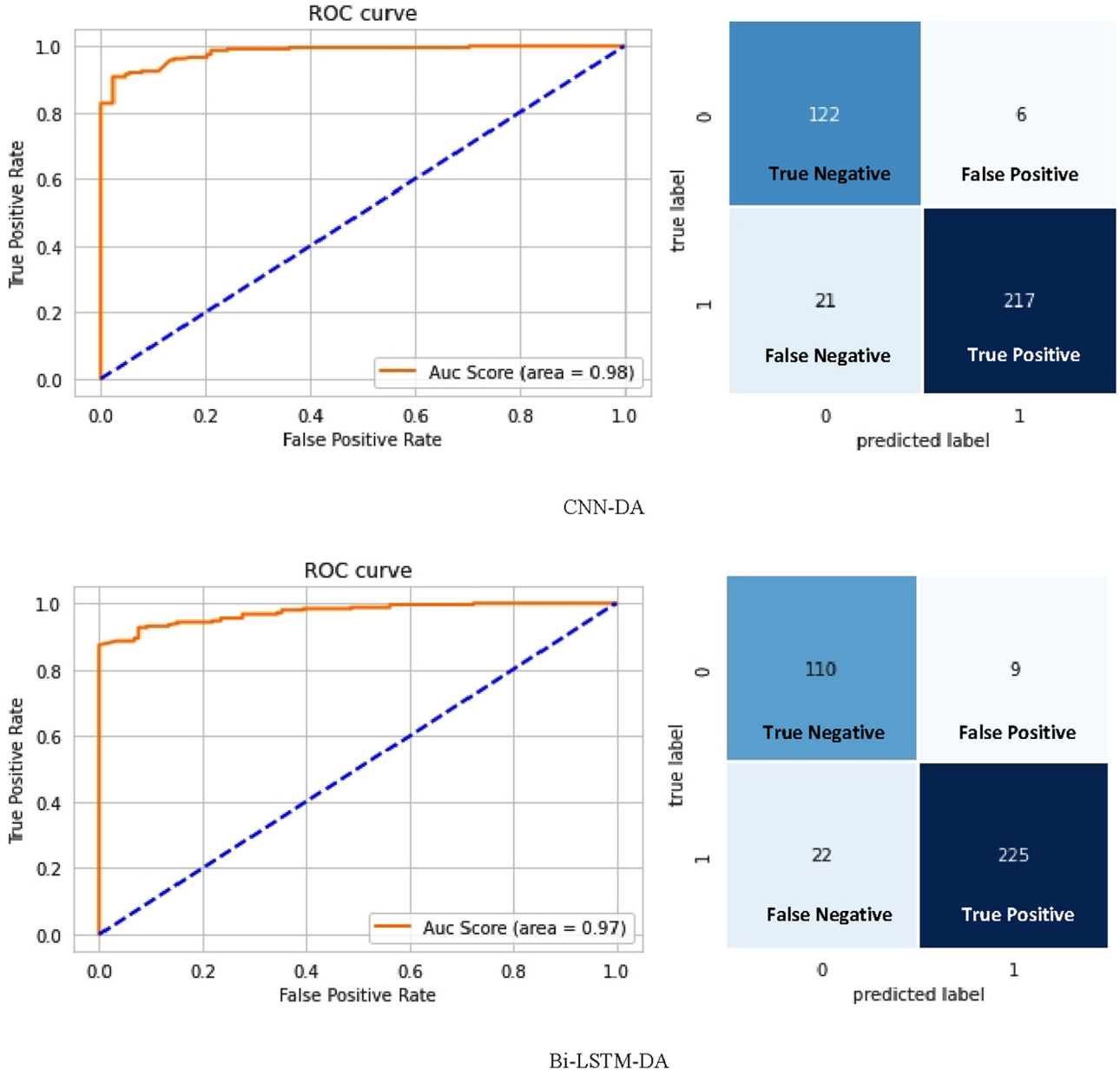


Figure 4. AUC and confusion matrix of the highest performing deep neural models with data augmentation. Note: There is ±0.05 change in the score of the plotted AUC.

techniques are applied to modify the distribution of the classes in a dataset to balance the number of positive and negative samples.[21] Swapping words approach was applied to the original dataset to help boost the signals of the not indicated samples (negative cases). It was noticed that there was a severe overfitting when the models trained on the original dataset, whereas using the augmented dataset helped eliminating the overfitting in both traditional and deep models (Table 1). Moreover, a significant improvement in the performances of the models has been noticed, with an increase in F1 scores ranging from 1% to 7%. Our results are in alignment with other NLP research that investigated the effects of data augmentation.[21–23]

Traditional models

With the augmented dataset, all models exhibited high AUCs and F1 scores. RF with BoW approaches exceeded other models with AUCs and F1 scores reaching 0.99 (Table 1). LR with TF-IDF demonstrated the highest results with AUC = 0.76 and F1 = 0.94 for the original dataset, suggesting that the greater the balance between positive and negative samples within a dataset, as in the augmented dataset, the better RF performs. BoW, bigrams, and trigrams with RF models slightly exceeded TF-IDF in the augmented dataset (Table 1). These results contradict the findings of Brown and Kachura,[24] in which the researchers reported lower performance for BoW approaches than TF-IDF. However, in the original dataset, models with TF-IDF exhibited similar or slightly higher performance than BoW approaches. One reason is that TF-IDF is a word frequency-dependent approach in which the weight of a word increases as its frequency in a single referral increases, and as the word frequently appears in all referrals in the dataset, its weight decreases and makes it difficult to distinguish the referrals. In the augmented dataset, we only augmented the not indicated requests using the words swapping technique, so each request was re-

Table 2
Classification results of the top 5 models and the eight radiographers on the held-out dataset. And the agreement between each classifications and the reference labels.

Models	country	Year of experience as MRI radiographer	Acc	Pre	Rec	F1	Kapp agreement with the reference labels of the 100 referrals	95% CI	Agreement level
RF–BoW–DA			0.84	0.84	0.97	0.90	0.31 (p = <0.001)	0.07–0.55	Fair
RF–Bigram–DA			0.83	0.83	0.98	0.90	0.20 (p = 0.004)	−0.01–0.43	Slight
RF–Trigram–DA			0.83	0.83	0.98	0.90	0.20 (p = 0.004)	−0.01–0.43	Slight
CNN–DA			0.83	0.86	0.93	0.90	0.35 (p = <0.001)	0.11–0.59	Fair
Bi-LSTM–DA			**0.86**	**0.92**	**0.90**	**0.91**	0.56 (p = <0.001)	0.35–0.76	Moderate
Radiographers									
Radiographer 1	Ireland	21	0.37	0.87	0.25	0.39	0.04 (p = 0.352)	−0.04–0.13	Slight
Radiographer 2	Ireland	20	0.67	0.92	0.64	0.75	0.28 (p = 0.001)	0.12–0.45	Fair
Radiographer 3	Malta	35	0.52	0.77	0.56	0.65	−0.08 (p = 0.354)	−0.2–0.08	No
Radiographer 4	Malta	21	0.57	0.88	0.54	0.67	0.14 (p = 0.074)	−0.01–0.29	Slight
Radiographer 5	UK	21	0.38	0.85	0.28	0.42	0.03 (p = 0.516)	−0.06–0.13	Slight
Radiographer 6	UK	7	0.80	0.88	0.87	0.88	0.39 (p = <0.001)	0.17–0.61	Fair
Reporting Radiographer 7	UK	30	0.28	1.00	0.11	0.19	0.04 (p = 0.128)	0.01–0.08	Slight
Reporting Radiographer 8	UK	20	0.79	0.89	0.83	0.86	0.38 (p = <0.001)	0.16–0.59	Fair

Note: Acc = accuracy, pre = precision, Rec = recall, F1 = F1 score, and CI = confidence interval. The bold values represent the highest performing model.

created 3 more times with the same words but different locations. As a result, the frequency of any given word in the minority class referrals increased in the augmented dataset and caused some incapability to TF-IDF to discriminate those words. Whereas in the original dataset, where no data augmentation was implemented, the frequencies of words that appears in the not indicated requests were low, which helped TF-IDF distinguish those words and exhibit higher classification performance than its performance in the augmented dataset. From these findings, we could indicate that with an imbalanced dataset, TF-IDF performs better than BoW, and when applying swapping words augmentation, the performance of TF-IDF deteriorates accordingly.

Deep neural models

CNN-DA slightly outperformed Bi-LSTM-DA and these findings are in line with.[25,26] This study shares more similarities with Chen et al.,[27] who built a CNN model using a pre-trained words embedding model, called Glove, to binary categorise computed tomography reports as positive or negative for pulmonary embolism. Their model achieved an AUC and F1 of 0.99 and 0.938, respectively, which is similar to our CNN–DA (AUC = 0.981, F1 = 0.94). However, their model was trained on a larger dataset with 2500 reports, whereas our model trained only on 854 referrals, and as it is known that deep neural models require more data to perform well.[26] The high classification results of our CNN–DA could be because of two things: (1) improving the balance between indicated and not indicated referrals within the augmented dataset and (2) integrating our model with Fasttext word embeddings. Fasttext is well suited to dealing with out of vocabulary (OOV) tokens (words), which are common in the medical texts. Fasttext divides each token in the dataset into n-gram characters and searches for similarities with words vectorisations in the vocabulary corpus that had been used to train Fasttext. Unlike the Glove model, which creates the vector of a whole token and if the token does not present in the vocabulary corpus that the Glove had been trained on, it turns as OOV, and hence impact the final prediction.[28,29]

Both models, CNN and Bi-LSTM, that developed using the original dataset, in which 10.8% of the LSMRI requests in the training set were not indicated, could not classify most of the not indicated referrals in the test set correctly. However, Bi-LSTM was able to accurately classify more not indicated referrals than CNN, 9 and 4, respectively. Although not satisfactory, these results suggest that Bi-LSTM might be more powerful than CNN in two-class classification tasks with imbalanced dataset, especially if the goal is to accurately classify the minority class.

Comparison with national and international MRI radiographers

In most of the available comparative research, comparisons between models and humans were conducted indirectly in which the agreement between the models and the reference labels were compared to the agreement between the annotators and the reference labels. However, this method does not give a clear comparison as the annotators were involved in establishing the reference labels that had been used for the development of the classifiers, and as a result, some bias may be occurred towards the annotators. In this study, the highest performing models were compared with eight national and international, reporting and non-reporting MRI radiographers, who were not participated in the creation of the reference labels, on a challenging held-out dataset (100 referrals; 81 indicated, 19 not indicated). Overall, the performances of all the highest models surpassed that of the radiographers in assigning the reference labels (Table 2). All models showed higher agreement with the reference labels of the 100 referrals than that demonstrated by the radiographers (Table 2). Bi-LSTM-DA achieved the highest F1 of 0.91, and classified 13 out of 19 not indicated referrals, and 73 out of 81 indicated referrals correctly. Whereas the highest MRI radiographer achieved F1 of 0.88, with correct classification of 10 out of 19 not indicated, and 70 out of 81 indicated referrals. It is important to acknowledge that some referrals can be confounders due to lack of, or misleading information. The purpose of this comparison, however, was not to prove the superiority of AI models over professionals but rather to demonstrate their efficacy in the medical area by comparing them with expert specialists. The findings indicate the capability of AI technologies in vetting LSMRI referrals with at least equivalent performance to healthcare professionals.

Analysis of misclassified referrals in the held-out dataset produced by the highest performing models

Manual analysis was performed for errors produced by the highest RF models when re-tested on the held-out dataset (Table 2). The main errors originated from referrals that included examinations for multiple regions/irrelevant information to the lumbar spine. Examples of false-positive predictions: "*previous cervical disc*

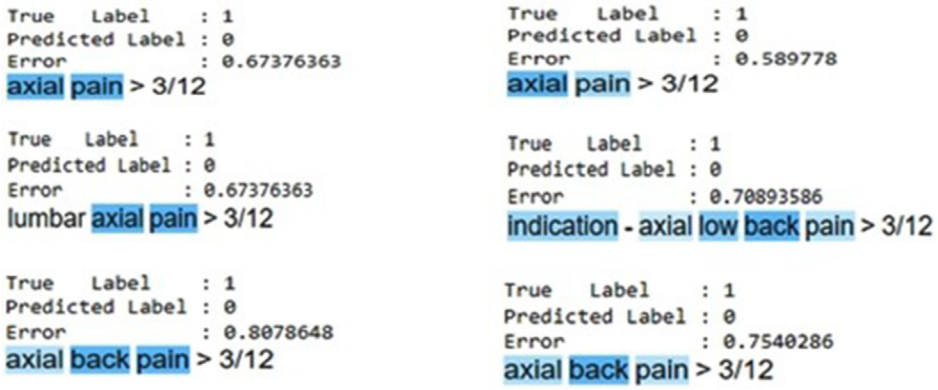


Figure 5. Image shows the sensitivity analysis (heatmap analysis) for three false-negative referrals that CNN–DA predicted (left) and three false-negative referrals by Bi-LSTM–DA (right).

replacement: recurrent neck and back pain", and "*lower pressure headache. Back pain at onset. ? Evidence of CSF leak?*", those referrals included information related to cervical disc and headache which presumably mislead the models. A similar errors source was also noticed in the false-negative predictions. As examples: "*Dx of breast Ca. Ongoing back pain. She has a cutaneous cyst,? cause*", and "*Abnormal neurologic examination of lower limbs and also Hoffman's positive both upper limbs*", both requests contained irrelated information to lumbar spine, which confused the classifiers, hence, the models incorrectly classified those referrals as not indicated.

Interpretation errors in DL models are challenging because such models are black boxes. However, to understand the error sources, a method called sensitivity analysis,[30] was implemented to display misclassified referrals in the shape of heatmap. This heatmap displays the documents and colours each word in the text with a shade of blue so that the word in dark blue indicates that this word was given higher weight by the model when categorising the document and vice versa. For false-negative predictions, it is obvious that the time codes were ignored in CNN-DA and Bi-LSTM-DA. Those codes have important meanings for determining the pain period and considered important indicators in assigning the referrals' appropriateness (Fig. 5). In false-positive predictions, most of the errors occurred in referrals with multiple examinations/irrelevant information to lumbar spine (Fig. 6).

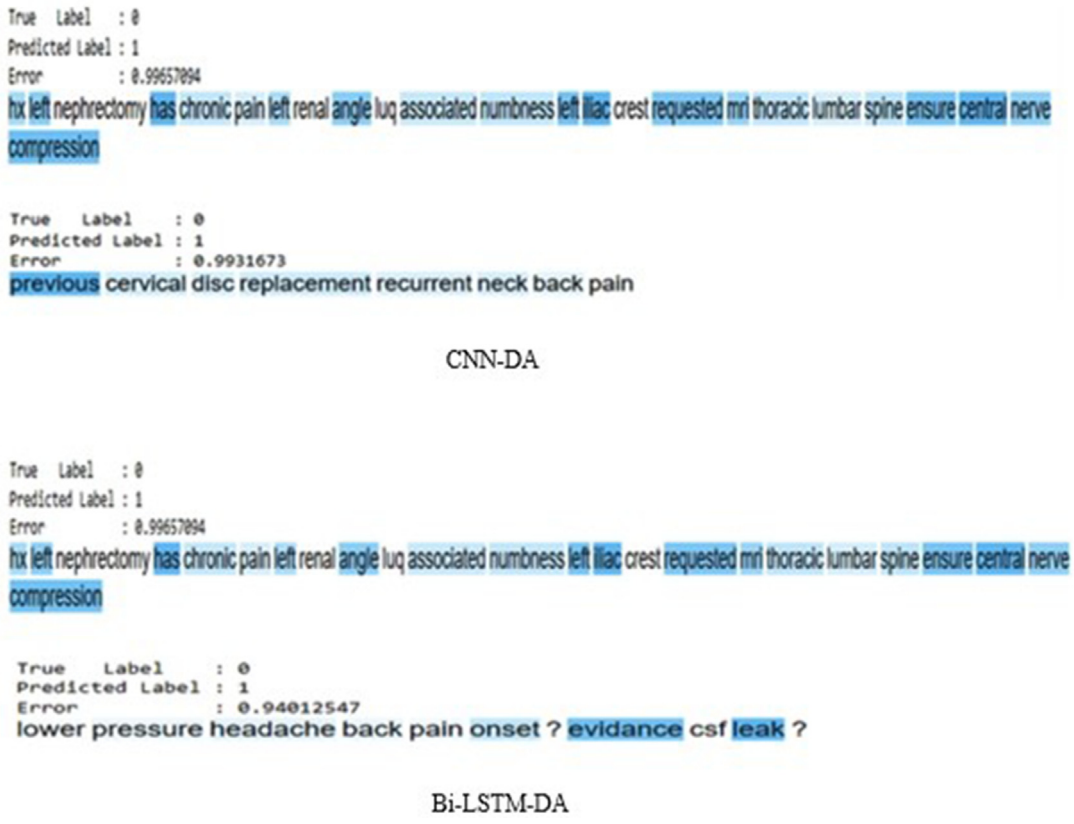


Figure 6. Image shows the sensitivity analysis (heatmap analysis) for two false-positive referrals that CNN–DA predicted (upper) and two false-positive referrals by Bi-LSTM–DA (lower).

Limitations

As discussed in the paper, a larger development dataset would improve model training and further research is warranted to incorporate other internationally applied referral guidelines and any variables in LSMRI guidance. Additionally, referrals with indications for multiple examinations confused the models and resulted in classification errors and this requires further consideration. Finally, we experimented with simple neural architecture and did incorporate advanced architectures such as attention-based recurrence neural network and multi-channel CNN, which might solve the issue of temporal codes ignorance we reported in the errors analysis.

Conclusion

We report the results of several traditional and deep neural classifiers developed for auto-vetting of LSMRI referrals. Data augmentation positively impacts models' performances, and models trained on the oversampled data demonstrated the ability to categorise the held-out referrals with equivalent or higher accuracy than reporting and non-reporting radiographers. These results suggest that machine learning applications for auto-vetting the referrals appropriateness are feasible and could improve the referrals' management.

Availability of data and materials

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Ethics approval and consent to participate

Ethical approval were obtained from the relevant institutional review from University College Dublin (Reference Numbers: LS-E-19-171-Alanazi-Rainford) and (LS-E-19-69-Alanazi-Rainford).

Authors' contributions

Ali Alanazi collected the data from the hospitals, developed the models, wrote the manuscript, and perform the analysis. Andrea Cradock was involved in the design, collection of the data, and text revision. Louise Rainford was involved in formulating the original idea, recruitment of participants, revised the manuscript critically.

Conflict of interest statement

The authors declare that they have no competing interests.

Acknowledgements

This research was funded by the ministry of education in the kingdom of Saudi Arabia [grant number KSP12014910].

References

1. Patel ND, Broderick DF, Burns J, Deshmukh TK, Fries IB, Harvey HB, et al. ACR appropriateness criteria low back pain. *J Am Coll Radiol* 2016;**13**(9):1069–78.
2. The Royal College of Radiologists. *RCR iRefer Guidelines: making the best use of clinical radiology*. London: The Royal College of Radiologists; 2017. Available at: https://www.rcr.ac.uk/clinical-radiology/being-consultant/rcr-referralguidelines/about-irefer. [Accessed 20 February 2022].
3. Kovacs FM, Arana E, Royuela A, Cabrera A, Casillas C, Pinero P, et al. Appropriateness of lumbar spine magnetic resonance imaging in Spain. *Eur J Radiol* 2013;**82**(6):1008–14.
4. Watura C, James S. Review of general practitioner direct access referrals for lumbar spine MRI. *Clin Radiol* 2013;**68**(2013):S5.
5. Avoundjian T, Gidwani R, Yao D, Lo J, Sinnott P, Thakur N. Evaluating two measures of lumbar spine MRI overuse: administrative data versus chart review. *J Am Coll Radiol* 2016;**13**(9):1057–66.
6. Flaherty S, Zepeda ED, Mortele K, Young GJ. Magnitude and financial implications of inappropriate diagnostic imaging for three common clinical conditions. *Int J Qual Health Care* 2019;**31**(9):1–7.
7. Baker R, Lecouturier J, Bond S. Explaining variation in GP referral rates for x-rays for back pain. *Implement Sci* 2006;**1**(15):1–6.
8. Kennedy SA, Fung W, Malik A, Farrokhyar F, Midia M. Effect of governmental intervention on appropriateness of lumbar MRI referrals: a canadian experience. *J Am Coll Radiol* 2014;**11**(8):802–7.
9. Wang KY, Yen CJ, Chen M, Variyam D, Acosta TU, Reed B, et al. Reducing inappropriate lumbar spine MRI for low back pain: radiology support, communication and alignment network. *J Am Coll Radiol* 2018;**15**(1):116–22.
10. Blackmore CC, Mecklenburg RS, Kaplan GS. Effectiveness of clinical decision support in controlling inappropriate imaging. *J Am Coll Radiol* 2011;**8**(1):19–25.
11. Liu C, Desai S, Krebs LD, Kirkland SW, Keto-Lambert D, Rowe BH. Effectiveness of interventions to decrease image ordering for low back pain presentations in the emergency department: a systematic review. *Acad Emerg Med* 2018;**25**(6):614–26.
12. Min A, Chan VW, Aristizabal R, Peramaki ER, Agulnik DB, Strydom N, et al. Clinical decision support decreases volume of imaging for low back pain in an urban emergency department. *J Am Coll Radiol* 2017;**14**(7):889–99.
13. Trivedi H, Mesterhazy J, Laguna B, Vu T, Sohn JH. Automatic determination of the need for intravenous contrast in musculoskeletal MRI examinations using IBM Watson's Natural Language Processing Algorithm. *J Digit Imaging* 2018;**31**(2):245–51.
14. Hassanpour S, Langlotz CP, Amrhein TJ, Befera NT, Lungren MP. Performance of a machine learning classifier of knee mri reports in two large academic radiology practices: a tool to estimate diagnostic yield. *Am J Roentgenol* 2017;**208**(4):750–3.
15. Zhang AY, Lam SS, Liu N, Pang Y, Chan LL, Tang PH. Development of a radiology decision support system for the classification of MRI brain scans. In: *Proc – 5th IEEE/ACM int conf big data comput appl technol BDCAT*; 2018. p. 107–15.
16. Mongan J, Moy L, Kahn CE. Checklist for artificial intelligence in medical imaging (CLAIM): a guide for authors and reviewers. *Radiol Artif Intell* 2020;**2**(2):1–6.
17. McHugh ML. Lessons in biostatistics interrater reliability : the kappa statistic. *Biochem Med* 2012;**22**(3):276–82.
18. Pons E, Braun LM, Hunink M, Kors JA. Natural Language processing in radiology: a systematic review. *Radiology* 2016;**279**(2):329–43.
19. He H, Garcia EA. Learning from imbalanced data. *Stud Comput Intell* 2009;**807**(9):81–110.
20. Khalilia M, Chakraborty S, Popescu M. Predicting disease risks from highly imbalanced data using random forest. *BMC Med Inform Decis Mak* 2011;**11**(1):1–13.
21. Abdollahi M, Gao X, Mei Y, Ghosh S, Li J. A dictionary-based oversampling approach to clinical document classification on small and imbalanced dataset. In: *IEEE/WIC/ACM int jt conf web intell intell agent technol WI-IAT*; 2020. p. 357–64.
22. Abulaish M, Sah AK. A text data augmentation approach for improving the performance of CNN. In: *11th int. conf commun syst networks, COMSNETS*; 2019. p. 625–30.
23. Wei J, Zou K. EDA: easy data augmentation techniques for boosting performance on text classification tasks. In: *9th international joint conference on natural language processing*; 2020. p. 6382–8.
24. Brown AD, Kachura JR. Natural language processing of radiology reports in patients with hepatocellular carcinoma to predict radiology resource utilization. *J Am Coll Radiol* 2019;**16**(6):840–4.
25. Heo TS, Kim YS, Choi JM, Jeong YS, Seo SY, Lee JH, et al. Prediction of stroke outcome using natural language processing-based machine learning of radiology report of brain MRI. *J Pers Med* 2020;**10**(4):1–11.
26. Dahl FA, Rama T, Hurlen P, Brekke P, Husby H, Gundersen T, et al. Neural classification of Norwegian radiology reports: using NLP to detect findings in CT-scans of children. *BMC Med Inform Decis Mak* 2021;**21**(1):1–8.
27. Chen MC, Ball RL, Yang L, Moradzadeh N, Chapman BE, Larson DB, et al. Deep learning to classify radiology free-text reports. *Radiology* 2018;**286**(3):845–52.
28. Adipradana R, Nayoga BP, Suryadi R, Suhartono D. Hoax analyzer for Indonesian news using RNNs with fasttext and glove embeddings. *Bull Electr Eng Inform* 2021;**10**(4):2130–6.
29. Khattak FK, Jeblee S, Pou-Prom C, Abdalla M, Meaney C, Rudzicz F. A survey of word embeddings for clinical text. *J Biomed Inform* 2019;**4**:1–18.
30. Arras L, Horn F, Montavon G, Müller KR, Samek W. Explaining predictions of non-linear classifiers in NLP. In: *Proceedings of the 1st workshop on representation learning for NLP*; 2016. p. 1–7.