PLOS ONE

RESEARCH ARTICLE

Supervised and unsupervised language modelling in Chest X-Ray radiological reports

Ignat Drozdov[1]*, **Daniel Forbes**[2], **Benjamin Szubert**[1], **Mark Hall**[3], **Chris Carlin**[4], **David J. Lowe**[2]

1 Bering Limited, London, United Kingdom, **2** Emergency Department, Queen Elizabeth University Hospital, Glasgow, Scotland, **3** Radiology Department, Queen Elizabeth University Hospital, Glasgow, Scotland, **4** Department of Respiratory Medicine. Queen Elizabeth University Hospital, Glasgow, Scotland

* idrozdov@beringresearch.com

Citation: Drozdov I, Forbes D, Szubert B, Hall M, Carlin C, Lowe DJ (2020) Supervised and unsupervised language modelling in Chest X-Ray radiological reports. PLoS ONE 15(3): e0229963. https://doi.org/10.1371/journal.pone.0229963

Editor: Ulas Bagci, University of Central Florida (UCF), UNITED STATES

Received: October 23, 2019

Accepted: February 17, 2020

Published: March 10, 2020

Data Availability Statement: Our internal radiological report dataset cannot be shared publicly due to patient confidentiality. MIMIC-CXR data set can be accessed through https://physionet.org/content/mimic-cxr/.

Funding: This work is supported by Bering Limited and the Industrial Center for AI Research in Digital diagnostics (iCAIRD) which is funded by the Data to Early Diagnosis and Precision Medicine strand of the government's Industrial Strategy Challenge Fund, managed and delivered by Innovate UK on behalf of UK Research and Innovation (UKRI)

Abstract

Chest radiography (CXR) is the most commonly used imaging modality and deep neural network (DNN) algorithms have shown promise in effective triage of normal and abnormal radiograms. Typically, DNNs require large quantities of expertly labelled training exemplars, which in clinical contexts is a major bottleneck to effective modelling, as both considerable clinical skill and time is required to produce high-quality ground truths. In this work we evaluate thirteen supervised classifiers using two large free-text corpora and demonstrate that bi-directional long short-term memory (BiLSTM) networks with attention mechanism effectively identify Normal, Abnormal, and Unclear CXR reports in internal (n = 965 manually-labelled reports, f1-score = 0.94) and external (n = 465 manually-labelled reports, f1-score = 0.90) testing sets using a relatively small number of expert-labelled training observations (n = 3,856 annotated reports). Furthermore, we introduce a general unsupervised approach that accurately distinguishes Normal and Abnormal CXR reports in a large unlabelled corpus. We anticipate that the results presented in this work can be used to automatically extract standardized clinical information from free-text CXR radiological reports, facilitating the training of clinical decision support systems for CXR triage.

Introduction

Chest radiography (CXR) is the most commonly used imaging modality, with over two billion procedures performed annually [1]. There is a general consensus that an Artificial Intelligence (AI)-supported reporting of CXR images could be a valuable adjunct to imaging interpretation, providing substantial benefit in many clinical contexts, from improved workflow prioritization and clinical decision support to large-scale screening and global population health initiatives [2–4]. Indeed, deep learning algorithms have been successfully applied to detect heterogeneous thoracic disease [3, 5], triage normal and abnormal radiographs [2], and identify specific pathologies such as pulmonary tuberculosis [6], pneumonia [7], and lung cancer [8].

Deep learning models require large quantities of expertly labelled training exemplars [9] and the well-established computer science mantra "Garbage In, Garbage Out" holds especially true in clinical applications of AI [10]. Whilst the gold-standard of image annotation remains direct application of expert knowledge, the sheer size of the required datasets makes this

[Project number 104690]. Views expressed are those of the authors and not necessarily those of Bering, the iCAIRD Consortium members, the NHS, Innovate UK or UKRI. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. ID and BS are employees of Bering Limited. The funder provided support in the form of salaries for authors ID and BS, but did not have any additional role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. The specific roles of these authors are articulated in the 'author contributions' section.

Competing interests: ID and BS are employees of Bering Limited. This does not alter our adherence to PLOS ONE policies on sharing data and materials.

endeavour impractical [11]. Therefore, Natural Language Processing (NLP) approaches offer an opportunity to automate the annotation of free-text reports [12]. For example, the Medical Language Extraction and Encoding (MedLEE) system relies on controlled vocabulary and grammatical rules to convert free text into a structured database [13]. PeFinder, an NLP system for pulmonary embolism classification, uses pre-defined lexical cues and context terms to achieve high sensitivity and positive predictive value [14]. Finally, NegEx, utilises hand-crafted regular expression rules to identify pertinent negatives from patient discharge summaries [15]. Nevertheless, applying text mining techniques to radiological reports, which may contain broken grammar and misspellings, poses a number of challenges due to extensive variability in linguistic ambiguity. Indeed, in the publicly-available ChestXray14 [16] imaging dataset, labels do not accurately reflect the visual content of the images, with positive predictive values of 10–30% lower than the values presented in the original documentation [11].

Neural network-driven modelling of radiological language has been proposed to supersede the hand-crafted rules and grammatical relations of the traditional rules-based algorithms [17]. Recently, a bi-directional long short-term memory (BiLSTM) network, which does not use any hand-engineered features, was demonstrated to perform favourably in a corpus of CXR reports (f1 = 0.87) [18]. Similarly, a supervised approach using a Recurrent Neural Network (RNN) with attention mechanism achieves high accuracy on expert-labelled CXR dataset (f1 = 0.93) [19]. Finally, Convolutional Neural Networks (CNNs) have been used to extract pulmonary embolism findings from thoracic computed tomography reports, outperforming state-of-the-art NLP systems (f1 = 0.94) [17].

Multi-label annotation of abnormal reports has been the primary aim of radiological language models [2, 4, 18, 19]. Nevertheless, practicalities of day-to-day clinical workflows suggest that the ability to identify 'normal' images and remove them from worklists would be anticipated to generate significant efficiency and cost savings [2, 20]. In addition, for clinicians reviewing images at the point of care, accurate triage of abnormal findings has potential safety, clinical outcome, and assistive (e.g. reduced cognitive overload) benefits [21, 22].

In this study we describe an approach to automatically extract standardized clinical information from free-text CXR radiological reports. More specifically, it is anticipated that accurate identification of Normal and Abnormal entities (irrespective of clinical sign or pathology) will facilitate training of AI-enabled triage systems at scale. We evaluate the utility of classical supervised machine learning techniques as well as state-of-the-art Long Short-Term Memory networks (LSTM) in the context of large corpora of free-text reports from Greater Glasgow and Clyde Health Board (n = 500,000) and the Beth Israel Deaconess Medical Center (MIMIC-CXR database [n = 227,835]) [23]. Additionally, we use ivis, an unsupervised Siamese Neural Network-based algorithm [24], which accurately classifies radiological reports and visualises document embeddings. Finally, we explore generalisability of machine learning techniques across European and North American radiological report corpora.

Materials and methods

Radiology reports and data preparation

Internal training, validation, and testing sets were produced using an in-house corpus of 500,000 deidentified CXR reports provided by NHS Greater Glasgow and Clyde (GGC) Safe-Haven. NHS GGC is the largest health board in Europe and delivers health care for 1.1 million patients with seven cute hospital sites. The reports cover the period between January 2007 and January 2019. The repository consists of text typed or dictated by the clinicians after radiograph analysis and does not contain clinician or patient identifiable information such as names, addresses or dates of birth. The reports had a minimum of 1 word and maximum of

380 words, with an average of 33.2 words and standard deviation of 20.5 words. On average, there were 4.8 sentences per report. Prior to analysis, reports were converted to lower case and lemmatized. Numbers, punctuation marks, special characters, and words that occurred in fewer than three documents were discarded. The final vocabulary contained 9,598 words.

A random sample of 5,000 reports was selected from the corpus for the purpose of creating expert-labelled training, validation, and testing sets. The reports were manually labelled by a clinical fellow (DF) with special interest in Radiology. The annotation schema included three classes–Normal, Abnormal, and Uncertain. The decision on the labelling was guided by the Fleischner Society Glossary of Terms for Thoracic Imaging [25]. A report was deemed to be Normal if it was explicitly stated as such in the free-form report and if there were no reported medical or surgical paraphernalia (e.g. pacemaker, sutures). An Abnormal label was assigned to reports with at least one documented radiological sign or presence of medical or surgical paraphernalia. If a report was normal for the patient (e.g. hyperinflated lungs in a patient with known Chronic Obstructive Pulmonary Disease), the report was still categorised as Abnormal. In cases where insufficient clinical information was provided to reliably label a report as Normal or Abnormal, a label of Uncertain was assigned. Reports that were either blank or inconclusive (e.g. "see above", "same as above") were excluded from the labelling exercise. All reports were labelled using the open source text annotation tool Doccano [26]. The final labelled corpus consisted of 4,821 reports.

The external testing set was drawn from 227,835 radiographic studies recorded within the MIMIC-CXR database [23]. A random sample of 500 reports was selected from the corpus for the purpose of creating an expert-labelled testing set. Pre-processing and annotation were performed as above. Following exclusion of inconclusive reports (e.g. reported only as "as above", "see above"), the final external testing corpus contained 465 reports. The reports had a minimum of 2 words and maximum of 118 words, with an average of 13.4 words and standard deviation of 16.2 words. On average, there were 2.9 sentences per report.

Supervised report classification

Expert-annotated reports were used to train three types of supervised classifiers: non-neural (i.e. classical machine learning algorithms), LSTM-based, and attention-based models (Transformers).

Non-neural classifiers. The labelled corpus, consisting of Normal, Abnormal, and Unclear reports, was converted to term frequency-inverse document frequency (tf-idf) matrix and reduced to 100 dimensions using Singular Value Decomposition (SVD). Subsequently, the transformed matrix was randomised into training (80%) and testing sets (20%) using a stratified split (Fig 1). Five supervised machine learning algorithms were evaluated on tf-idf/SVD-transformed radiology reports–K-Nearest Neighbour Classifier (KNN), Logistic Regression (LR), Gaussian Naïve Bayes Classifier (NBC), Random Forest (RF), and Support Vector Machine (SVM). Each model's hyperparameters were tuned on the training set using a Grid Search algorithm with stratified five-fold cross-validation. Hyperparameters that yielded the best macro-averaged f1 statistic across the five folds were retained for predictions on the independent testing set.

LSTM classifiers. The internal labelled corpus, consisting of Normal, Abnormal, and Unclear reports, was randomised into training (80%) and testing sets (20%) using stratified splits. Each report was then represented as a tokenised sequence of words. We limited the maximum length of the input sequence to 40, padding shorter sequences with zeros, whilst cropping longer sequences. Model inputs were mapped to an Embedding layer, which was initialised either by using either pre-trained fastText [27] weights or by drawing from a uniform

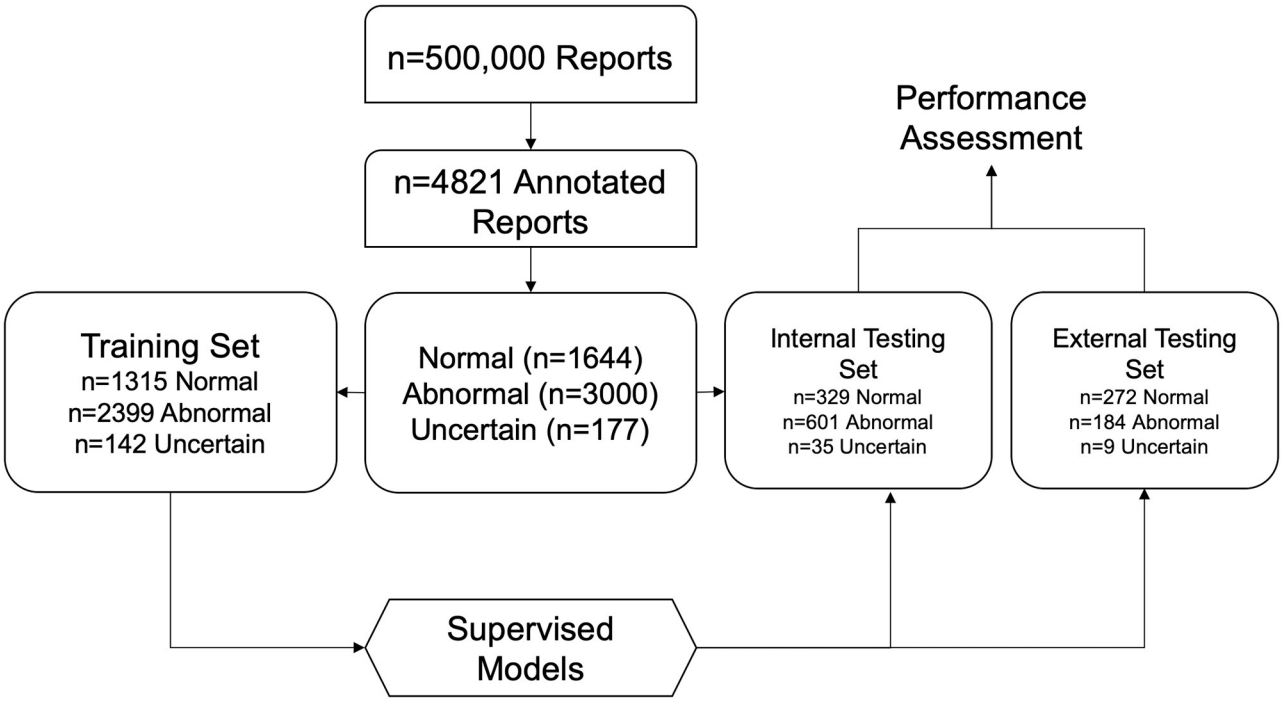

Fig 1. Flowchart showing supervised approach to radiology report classification.

https://doi.org/10.1371/journal.pone.0229963.g001

distribution in the (−0.01, 0.01) range. The fastText model was trained on an unlabelled corpus of lemmatised and pre-processed free-text reports (n = 495,179, see above). Window size was set to three and embedding dimensionality was set to 50. Subsequently, a Bidirectional LSTM (BiLSTM) architecture [28] was implemented, with each LSTM layer consisting of 100 memory cells. The loss function was the categorical cross-entropy between the predicted probabilities of the report tags and the true tags.

Our BiLSTM model was also supplemented with an attention mechanism (BiLSTM-ATT) [19, 29], in which the BiLSTM layer is followed with an attention module. The attention module generates a predictive distribution over the LSTM encodings for each step by firstly calculating the dot-product of the latest hidden state and the previous states, and then using the SoftMax function [30]. Applying these scores to the previous hidden state vectors effectively samples the most useful input vectors dynamically by predicting which vectors are most important for the predictions. By enabling selective sampling of relevant information from all encoder states, the model is able to deal with long sequences of words and maintain global information about the input sentence. Finally, all models were trained for 20 epochs with batches of 32 sentences using the Adam optimiser with the learning rate set to 0.001. Training was terminated early if the validation loss did not improve for three consecutive epochs.

Transformer classifiers. The Transformer is a novel neural network architecture based solely on a self-attention mechanism [31]. Four Transformer models were trained on the internal training set—Bidirectional Encoder Representations from Transformers (BERT) [32], DistilBERT [33], XLNet [34], and RoBERTa [35]. Each Transformer model was initialised using pre-trained weights provided by the HuggingFace's Transformers library [36]. A sequence-classification head (a linear layer) was added on top of the base model's hidden states. Radiology reports were then represented as a tokenized sequence according to the requirements of each of the Transformer models–using a punctuation and wordpiece tokenizer (BERT,

DistilBERT), SentencePiece tokenizer (XLNet), or Byte-Pair Encoding (RoBERTa). We limited the maximum length of the input sequence to 128, padding shorter sequences with zeros, whilst cropping longer sequences. Subsequently, all models were trained using the Adam optimizer for 20 epochs. Training was terminated early if validation loss did not improve for three consecutive epochs.

Unsupervised report classification

Reliable ground truths in radiological data is a scarce resource, which requires considerable clinical time and expertise. To address this limitation, we introduce a fully unsupervised approach to assigning Normal and Abnormal labels to free-text radiological reports.

Dimensionality reduction using siamese neural networks. The unsupervised ivis algorithm [24] was used to reduce dimensionality of 50-dimensional fastText embeddings of unlabelled reports within the GGC corpus. To obtain report-level embeddings, fastText word vectors within each report were averaged [37] and the resulting 50-dimensional vector was used to as inputs into the ivis algorithm. ivis Siamese Neural Network was initialised using three identical three-layer dense networks consisting of 500, 500, and 2,000 neurons each, followed by an embedding layer with the number of neurons reflecting dimensionality of desired embeddings. The layers preceding the embedding layer use the SELU activation function, which gives the network a self-normalizing property [38]. The weights for these layers are randomly initialized with the LeCun normal distribution. The embedding layers use a linear activation and have their weights initialized using Glorot's uniform distribution. The network was trained using a triplet loss function, whilst Euclidean distance was used to establish similarity between points in the embedding space [24]. Nearest neighbour selection was limited to 130 points and the training was halted early if the triplet loss did not improve for five epochs.

Gaussian mixture model clustering. A Gaussian Mixture Model (GMM) with two mixture components was applied to either FastText or ivis embeddings (Fig 2). Posterior probabilities of each mixture component were then obtained on the expertly labelled internal (GGC) and external (MIMIC-CXR) testing sets. The GMM's performance was evaluated by comparing ground truth labels of the testing set to mixture component probabilities.

CheXpert labeller. The CheXpert labeller is an NLP tool based on keyword matching with hardcoded rules describing negation [4], which assigns each report with one or more labels associated with thoracic pathology. The labeller operates in three stages: 1) extraction, 2) classification, and 3) aggregation. In the extraction stage, all mentions of a label are identified, including alternate spellings, synonyms, and abbreviations. Mentions are then classified as positive, uncertain, or negative using local context. In cases where keyword matching fails to produce a reliable result, a label of No Findings is assigned. We considered all reports with a label of No Findings to be Normal, whilst remaining reports were considered to be Abnormal.

Performance assessment

Model performance was assessed on internal (NHS GGC) and external (MIMIC-CXR Database) testing sets. The following performance metrics were recorded–precision, recall, f1-score, and Area Under Receiver Operating Characteristic Curve (AUROC). In multi-class classification problems, we weigh the average of the precision, recall, and f1-score by the number of instances of each class.

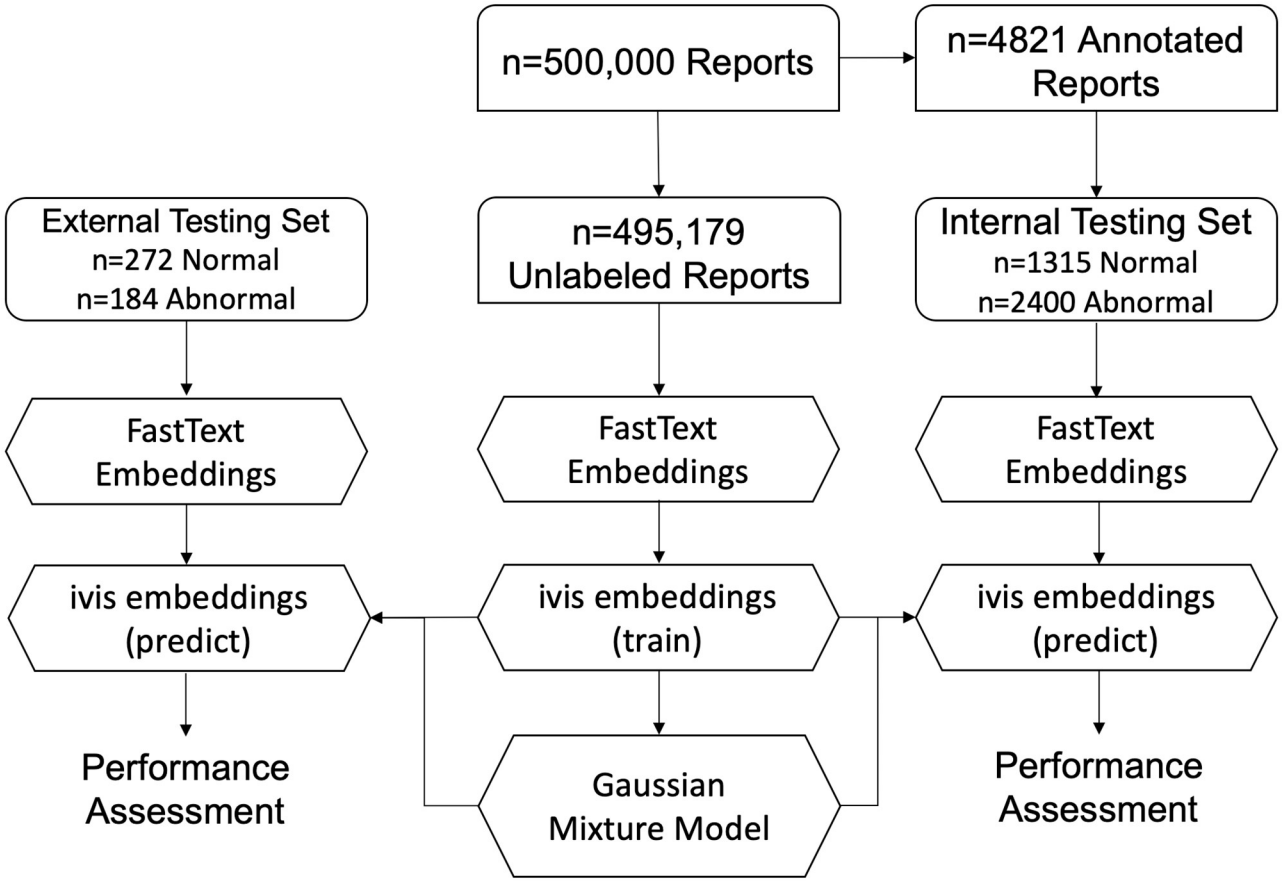

Fig 2. Flowchart demonstrating unsupervised approach to radiology report classification.

https://doi.org/10.1371/journal.pone.0229963.g002

Results

Supervised report classification

Five supervised multi-class classifiers were trained on tf-idf/SVD-transformed document matrices (n = 1,315 Normal, n = 2,399 Abnormal, n = 142 Uncertain)–KNN, Logistic Regression, Naïve Bayes, Random Forest, and SVM (Fig 3). For each model, an exhaustive grid search was carried out on the training set using 5-fold cross validation, optimising the f1 score, and the best performing parameters were fixed for subsequent performance assessment. SVM and Logistic Regression performed consistently well in identifying Normal and Abnormal reports, both in internal (AUROC 0.97–0.98, Fig 3B and 3E) and external (AUROC 0.97–0.98, Fig 3G and 3J) testing sets. Although SVM performed well in differentiating Unclear reports in the internal testing set (AUC = 0.86), all classifiers yielded suboptimal accuracy for this class in the external set (AUROC 0.39–0.51, Fig 3F–3J).

Next, we used the expert-labelled internal radiological reports to train a series of three-class BiLSTM classifiers using tokenised report sequences. We hypothesised that by considering temporal word relationships within each report, a more nuanced and generalisable model could be obtained through modelling radiological language with BiLSTMs. As above, performance was assessed on both internal and external testing sets. Pre-training BiLSTM with fastText embeddings (BiLSTM-fastText) produced robust classifiers compared to randomly initialised model weights (Fig 4A–4D, Table 1). Whilst BiLSTM-fastText resulted in marginally

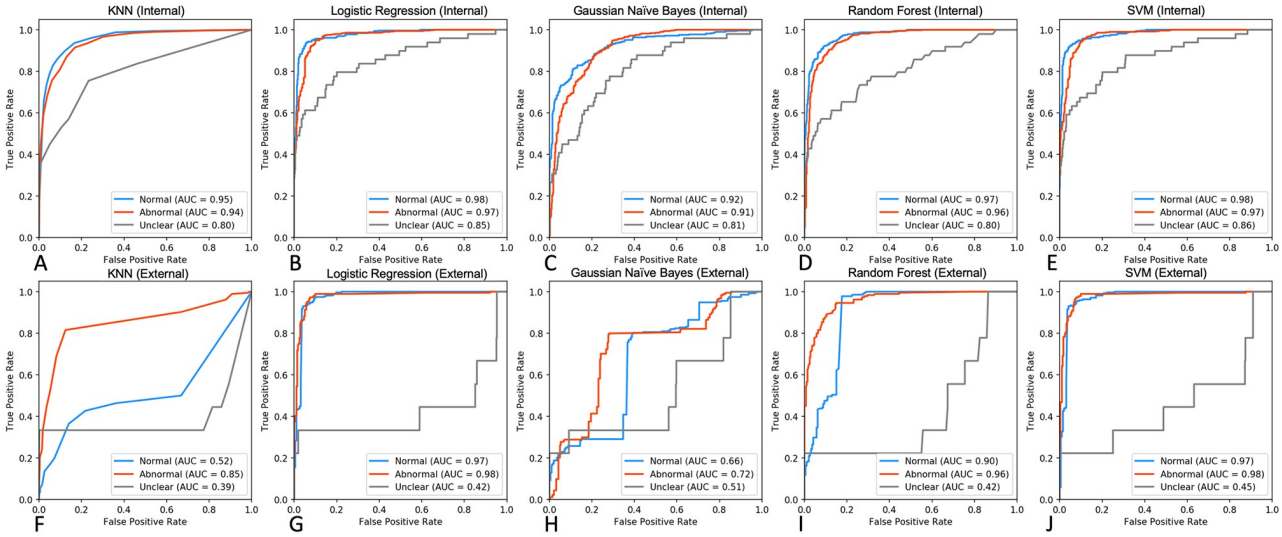

Fig 3. Performance assessment of non-neural classifiers on internal and external testing sets. A-E. ROC curves displaying performance metrics on an expert-labelled internal testing set (n = 329 Normal, n = 601 Abnormal, n = 35 Uncertain). **F-G.** ROC curves demonstrating classifier performance on external MIMIC-CXR free-text reports (n = 272 Normal, n = 184 Abnormal, n = 9 Uncertain).

https://doi.org/10.1371/journal.pone.0229963.g003

better detection of Unclear class compared to the top-performing SVM classifier ($AUC_{BiLSTM\text{-}fastTet}$ = 0.87 vs. AUC_{SVM} = 0.86, Fig 4A), this performance was further superseded by introducing attention mechanism to BiLSTM-fastText architecture ($AUC_{BiLSTM\text{-}Att\text{-}fastText}$ = 0.88, Fig 4C, Table 1).

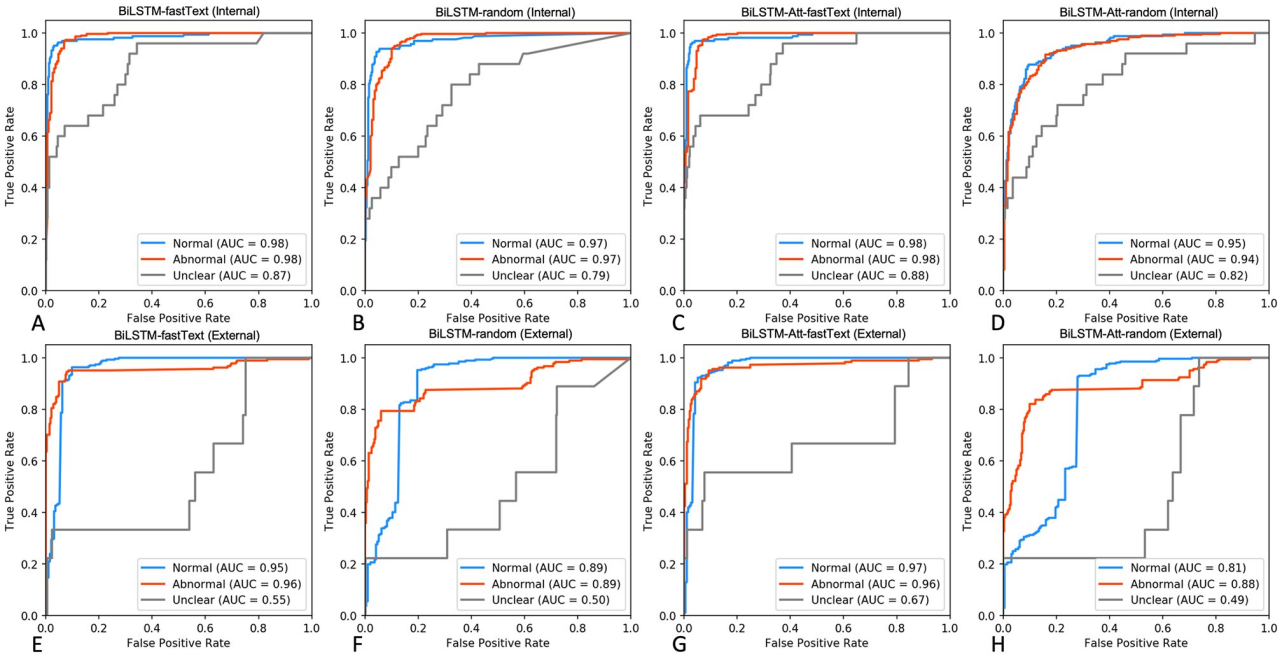

Fig 4. Performance assessment of BiLSTM classifiers on internal and external testing sets. A-D. ROC curves displaying performance metrics on an expert-labelled internal testing set (n = 329 Normal, n = 601 Abnormal, n = 35 Uncertain). **E-H.** ROC curves demonstrating classifier performance on the external MIMIC-CXR expert-labelled free-text reports (n = 272 Normal, n = 184 Abnormal, n = 9 Uncertain).

https://doi.org/10.1371/journal.pone.0229963.g004

Table 1. Performance comparison of supervised multi-class classifiers on internal and external testing sets. Class-weighted values are reported.

Classifier	Internal Testing Set (n = 978)			External Testing Set (n = 465)		
	Precision	Recall	F1-score	Precision	Recall	f1-score
KNN	0.82	0.86	0.84	0.81	0.82	0.81
Logistic Regression	0.85	0.90	0.87	0.91	0.93	0.92
Naïve Bayes	0.78	0.82	0.80	0.70	0.47	0.38
Random Forest	0.83	0.88	0.86	0.84	0.80	0.80
SVM	0.85	0.90	0.88	0.91	0.93	0.92
BiLSTM-fastText	0.93	0.93	0.93	0.91	0.91	0.91
BiLSTM-random	0.91	0.91	0.91	0.72	0.58	0.57
BiLSTM-Att-fastText	**0.94**	**0.94**	**0.94**	0.90	0.91	0.90
BiLSTM-Att-random	0.90	0.91	0.90	0.73	0.55	0.53
BERT	0.92	0.93	0.92	0.94	0.93	0.93
DistilBERT	0.91	0.91	0.91	0.93	0.93	0.93
XLNet	0.93	0.93	0.93	**0.95**	**0.95**	**0.95**
RoBERTa	0.91	0.91	0.91	0.93	0.93	0.93

https://doi.org/10.1371/journal.pone.0229963.t001

To assess how well our BiLSTM models generalise to an external testing set, we compared predicted labels to manually annotated reports from the MIMIC-CXR database (n = 272 Normal, n = 184 Abnormal, n = 9 Unclear, Fig 4E and 4F). Randomly-initialised BiLSTMs (BiLSTM-random) exhibited worse performance compared to both fastText-pretrained models and non-neural classifiers (Fig 4F and 4H). Interestingly, BiLSTM-Att-fastText, generalised well across Normal and Abnormal classes, and performed favourably on the Unclear class in the internal and external testing sets (Fig 4C and 4G).

Finally, four self-attention based models (BERT, DistilBERT, XLNet, and RoBERTa) were evaluated on the internal and external testing sets (Fig 5). Whilst all models performed well in classifying Normal and Abnormal reports (AUC = 0.97–0.99), XLNet achieved favourable

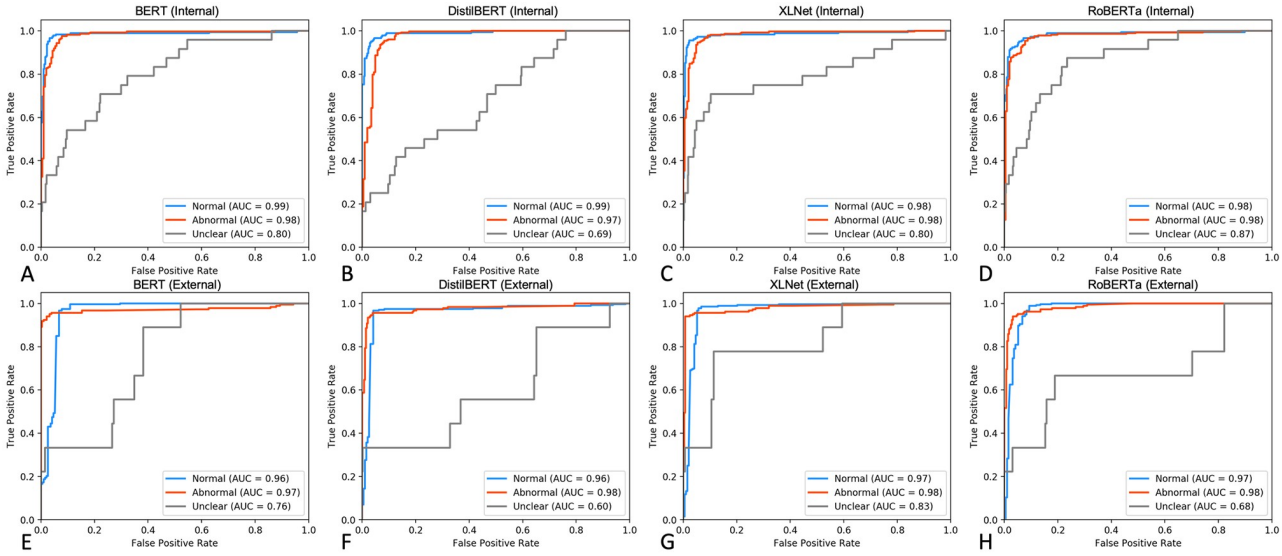

Fig 5. Performance assessment of Transformer-based classifiers on internal and external testing sets. A-D. ROC curves displaying performance metrics on an expert-labelled internal testing set (n = 329 Normal, n = 601 Abnormal, n = 35 Uncertain). **E-H.** ROC curves demonstrating classifier performance on the external MIMIC-CXR expert-labelled free-text reports (n = 272 Normal, n = 184 Abnormal, n = 9 Uncertain).

https://doi.org/10.1371/journal.pone.0229963.g005

performance on the Unclear class in both internal and external testing sets (AUC = 0.80 and AUC = 0.83 respectively, Fig 5C and 5G).

Unsupervised report classification

We demonstrated that supervised classifiers achieve excellent performance using a relatively small number of labelled training exemplars (n = 3,856 reports). Furthermore, neural networks that utilise the BiLSTM architecture with attention mechanism appear to generalise well to external radiological reports. Nevertheless, generation of reliable ground truths remains a barrier to training effective deep learning models due to the required clinical time and expertise. To address this limitation, we set out to develop and evaluate an unsupervised approach to assigning Normal and Abnormal labels to free-text radiological reports.

An internal corpus of n = 495,179 unlabelled reports was represented as a collection of 50-dimensional fastText document vectors (see Methods). We hypothesised that free-text entities can be modelled using Gaussian Mixture Distributions due to inherently distinct semantic structure of Normal and Abnormal reports. To test this hypothesis, a GMM with two components was constructed from the unlabelled document vectors and posterior probabilities of each component were extracted from the expert-labelled reports of the internal (n = 1,315 Normal and n = 2,400 Abnormal) and external (MIMIC-CXR, n = 272 Normal and n = 184 Abnormal) corpora. Whilst GMM performance on an internal testing set was sub-optimal, model validation on an external testing set produced acceptable metrics (AUC = 0.50 and AUC = 0.86 respectively, Fig 6D).

Recently, we introduced a novel algorithm, ivis, for dimensionality reduction and feature engineering in large datasets [14]. ivis is a parametric method that utilises a Siamese Neural Network to generate low-dimensional data representations that preserve both local and global properties of original observations. To further refine fastText embeddings, we applied ivis to 50-dimensional report vectors prior to GMM clustering (Fig 6A). Reduction of fastText reports to two-dimensional ivis representations resulted in marked performance improvements in both internal and external datasets (AUC = 0.89 and AUC = 0.90 respectively, Fig 6B and 6C). GMM performance was enriched further by expanding ivis representations to ten embedding dimensions (ivis-10D, $AUC_{Internal} = 0.89$, $AUC_{External} = 0.94$, Fig 6E and 6F).

Finally, we compared GMM-clustered ivis embeddings to annotations generated by the CheXpert Labeller. The labeller is a rule-based classifier which operates in three stages: 1) extraction, 2) classification, and 3) aggregation. In the extraction stage, all mentions of a label are identified, including alternate spellings, synonyms, and abbreviations. Mentions are then classified as positive, uncertain, or negative using local context. The CheXpert Labeller is tailored for CXRs, and recently demonstrated favourable performance on free-text reports [23]. Both the Labeller and GMM-clustered ivis-10D embeddings achieved comparable performances on an internal and external dataset ($f1\text{-}score_{Internal} = 0.81$ and $f1\text{-}score_{External} = 0.92$, Table 2). Interestingly, just two-dimensional ivis embeddings achieved acceptable classification performance, making the datasets amenable to interpretable visualisation (Fig 6A–6C).

Discussion

In this work we examine the application of supervised machine learning algorithms to classification of free-text CXR reports. Rigorous performance benchmarking on two independent corpora from two international health systems demonstrate that BiLSTM networks with self-attention mechanism produce state-of-the-art classification results and are generalise to external testing sets. Furthermore, we introduce a fully unsupervised approach for abnormality

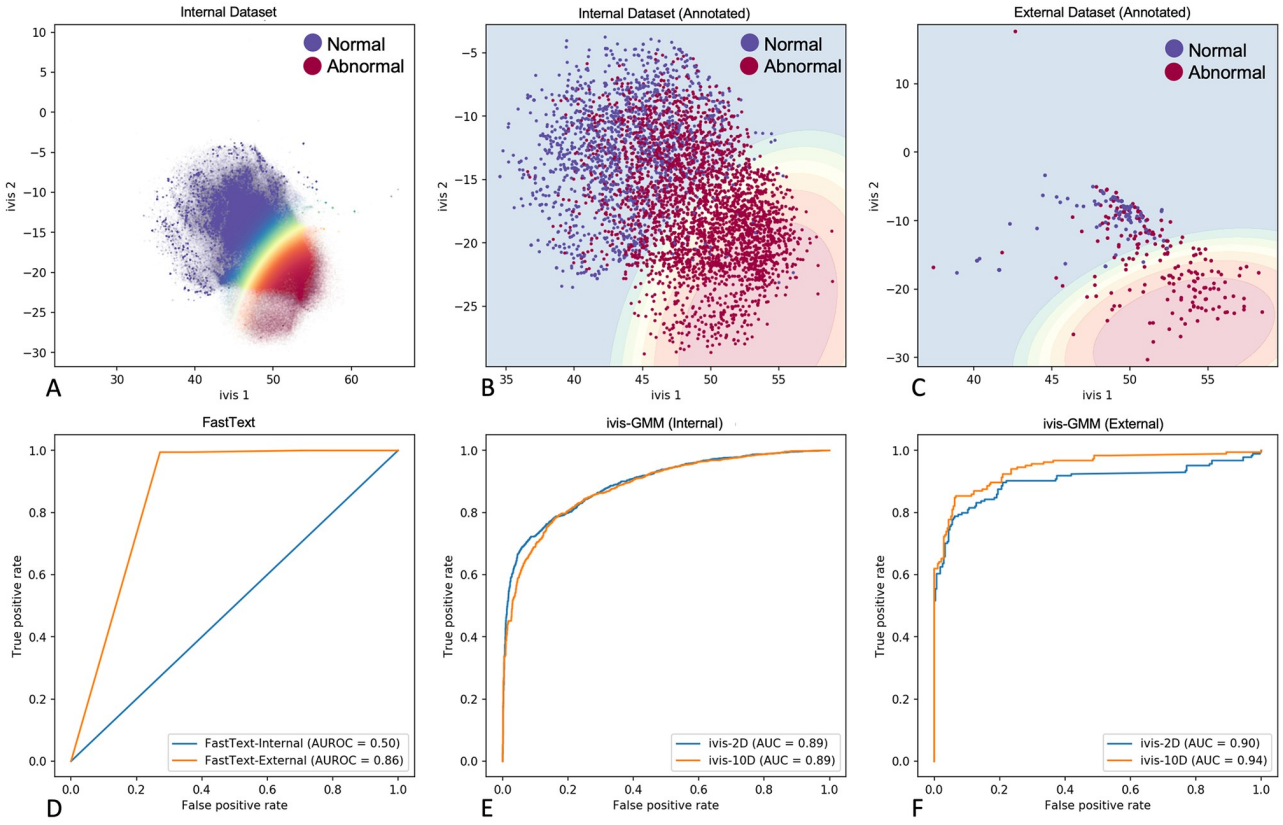

Fig 6. Unsupervised report classification using fastText, ivis, and Gaussian Mixture Model clustering. A. Two-dimensional ivis representation of 50-dimensional fastText embeddings of n = 495,179 unlabelled radiological reports from NHS GGC. Colour gradient reflects posterior probability of Normal and Abnormal report cluster. **B.** Scatterplot of predicted ivis embeddings for n = 3,715 expert-labelled reports in the internal testing set. Blue and red points represent manually-labelled Normal and Abnormal reports respectively. Colour gradient reflects contours of posterior probability distributions obtained from GMM model trained on two-dimensional ivis representations of n = 495,179 unlabelled radiological reports. **C.** Scatterplot of predicted ivis embeddings for n = 456 expert-labelled reports in the MIMIC-CXR testing set. Blue and red points represent manually-labelled Normal and Abnormal reports respectively. Colour gradient reflects contours of posterior probability distributions obtained from GMM model trained on two-dimensional ivis representations of n = 495,179 unlabelled radiological reports. **D**. ROC curves of unsupervised GMM classifier applied to 50-dimensional fastText embeddings of internal (n = 3,715) and external (n = 456) manually-labelled reports. **E-F**. ROC curves of unsupervised GMM classifier applied to two- and ten-dimensional ivis embeddings of manually labelled internal (n = 3,715) and external (n = 456) reports.

https://doi.org/10.1371/journal.pone.0229963.g006

detection in free-text reports, which performs favourably compared to a well-established rules-based classifier tuned for CXR labelling.

Our analysis of five non-neural supervised classifiers (KNN, Logistic Regression, Naïve Bayes, Random Forest, and SVM) demonstrated that whilst all models achieved excellent performance on an internal testing set, only SVM successfully captured Normal and Abnormal entities in both testing sets. Reports labelled as Unclear were consistently misclassified by all

Table 2. Performance comparison of unsupervised classifiers on internal and external radiological reports. Average performance values are reported.

Classifier	Internal Testing Set (n = 3715)			External Testing Set (n = 456)		
	Precision	Recall	f1-score	Precision	Recall	f1-score
fastText+GMM	0.42	0.65	0.51	0.88	0.84	0.84
ivis-2D+GMM	**0.83**	0.80	0.80	0.88	0.88	0.88
ivis-10D+GMM	0.82	0.80	**0.81**	0.91	0.91	**0.92**
CheXpert Labeller	0.81	**0.81**	**0.81**	**0.93**	**0.93**	**0.92**

https://doi.org/10.1371/journal.pone.0229963.t002

algorithms in the external testing set (Fig 3F–3J). These findings are consistent with the general notion that SVMs are well-suited for a text classification task due to 1) the algorithm's ability to learn independently of the dimensionality of the feature space, 2) suitability to problems with dense concepts and sparse instances (document vectors are sparse as each document vector contains only few entries which are not zero), and 3) linearly separable nature of most text problems [39]. Indeed, an SVM trained on a bag of phrases was used to detect hospital admissions due to specific diseases [40] as well as classify medical subdomain across clinical notes [41]. Interestingly, the SVM only marginally outperforms Logistic Regression classifier. Considering that Logistic Regression predictions may be viewed as locally interpretable [42], the minor trade-off in accuracy may be justified in favour of trusting and understanding intuition behind each classification [43].

Non-neural classifiers rely on a bag-of-words (BOW) representation of the training corpus. The approach maintains word multiplicity, but disregards grammatical nuances and word order of original sentences. Additionally, BOW matrices are often sparse, with only a few entries which are not zero. This convention has often proven to be problematic for non-neural classifiers due to the data sparsity problem [44]. In recent years, deep artificial neural networks have been found to yield consistently good and often state-of-the-art results on a variety of NLP tasks [18]. It can be argued that by considering complex inter-relationships between words within sentences, deep neural networks achieve state-of-the-art performance across NLP tasks such as part-of-speech tagging, shallow parsing, named entity recognition, and semantic role labelling [45].

We demonstrated that BiLSTM networks learn to differentiate Normal and Abnormal CXR reports and generalise well to an independent testing set (Table 1). Traditionally, BiLSTMs have shown performance improvements in NLP tasks over Unidirectional LSTMs, lkely due to inclusion of information from both future and past words in the sentence [28]. We demonstrate that an important requirement to a generalisable model is initialising the network with pre-trained word embeddings. Indeed, pre-trained BiLSTMs weights considerably outperformed random weight initialisation in terms of precision, recall, and f1-scores (Table 1). Previous reports have shown only marginal accuracy gains attributed to pre-training [18]. However, this is like because only an internal testing set was used to benchmark algorithm performance, whilst we note considerable gains on external datasets.

To pre-train our models, we applied fastText to an unlabelled corpus of n = 495,179 reports from NHS GGC. Several important features prompted us to choose fastText over other comparable approaches. First, the algorithm is fast and can train on our corpus within a few minutes. This allowed us to experiment with hyperparameters in order to produce better embeddings. Second, fastText operates at a character level, meaning that word vectors can still be extracted for those words that are not present in the original vocabulary. This is especially important as spelling and abbreviations vary greatly between radiological reports [46]. Indeed, it is not unreasonable to hypothesise that this feature of the fastText algorithm contributed significantly to model generalisability across testing sets. Finally, unlike word vectors from word2vec [47], fastText word features can be averaged together to form good sentence representations [37]. It is plausible that adding more reports into the pre-training corpus will lead to further performance gains–this is something that we intend to explore in greater detail as our work evolves.

In an earlier work, the attention mechanism was demonstrated to achieve high accuracy on an expertly labelled CXR dataset (f1-score = 0.93) [19]. The attention layer learns heterogeneous text representations for each label under an assumption that each snippet containing distinguishing information could be anywhere in the text and would differ across labels [19]. As such, attention can help combine global and local information in order to improve

classification performance [48]. In this work, a BiLSTM (pre-trained with fastText vectors) with attention mechanism was the top-performing LSTM classifier (f1-score$_{Internal}$ = 0.94, f1-score$_{External}$ = 0.90, Table 1). Interestingly, the approach also identified Uncertain reports considerably better on an external testing set (AUC = 0.67, Fig 4G), suggesting that more nuanced information can still be learnt from small number of exemplars. Overall, this work, together with independent reports [19, 48, 49], suggests that training a BiLSTM with attention on a relatively small corpus of labelled data produces generalisable state-of-the-art free-text classifiers that may augment training of computer vision models.

Recently, several novel network architectures, based solely on attention mechanisms, have achieved state-of-the-art performance across NLP tasks [31]. In this work we demonstrate that four Transformer-based models, namely BERT, DistilBERT, XLNet, and RoBERTa, achieve excellent performance (AUC: 0.97–0.99, Fig 5) on free-text radiological reports, which generalises well to an external testing set. Although performances of all Transformer-based models were comparable to BiLSTM with attention mechanism for Normal and Abnormal reports, XLNet identified Uncertain reports with increased accuracy (AUC$_{XLNet}$ = 0.80–0.83 vs. AUC$_{BiLSTM\text{-}Att\text{-}fastText}$ = 0.67–0.88). This improvement is likely due to the capacity of XLNet to learn bidirectional contexts and its autoregressive formulation [34]. Nevertheless, despite marginal increase in performance, training and finetuning Transformer-based models is a computationally expensive task. Marginal performance gains are offset by the hardware resources required to complete training and inference experiments.

So far, we have shown that supervised classifiers achieve excellent performance using a relatively small number of labelled training exemplars (n = 3,856 reports). Nevertheless, generation of reliable ground truths remains a barrier to training effective deep learning models due to the required clinical time and expertise [11]. To address this limitation, we set out to develop and evaluate an unsupervised approach to assigning Normal and Abnormal labels to free-text radiological reports. Our top-performing approach involves three steps: 1) obtaining sentence vectors by averaging fastText word features, 2) feature extraction using ivis, a novel Siamese Network algorithm, and 3) fitting a GMM with two components (assuming that distributions of Normal and Abnormal reports are inherently different) to ivis embeddings. Performance of this three-step approach was comparable to the CheXpert Labeller, which utilises hand-crafted rules tailored for CXR report annotation [4, 23]. We have previously applied ivis to structured single-cell datasets [24], demonstrating that the algorithm reliably preserves local and global distances in a low-dimensional space. Briefly, ivis employs a Siamese Neural Network architecture that learns to discriminate between similar and dissimilar fastText vectors without imposing strong priors. This property enables natural creation of dense clusters with shared nearest neighbours, making sentence vectors amenable to modelling with GMMs. Interestingly, although ivis was trained on the unlabelled internal corpus, it performed considerably better on an external testing set (Table 2). This was also the case for the CheXpert Labeller. It is likely that given that external testing set reports were shorter than internal reports (external average: 13.4 words vs. internal average: 33.2 words), the external reports were more linearly separable (Fig 6C), resulting in improved unsupervised performance.

Whilst ivis+GMM performance was comparable to CheXpert Labeller, we have identified several advantages of our unsupervised approach. First, GMMs can be used to obtain posterior probabilities of each component for every report. This provides a degree of granularity to our results. For example, at a component probability threshold greater than 0.99, ivis+GMM identified 30% of Abnormal reports with 100% positive predictive value. Conversely, the CheXpert Labeller provides strictly categorical outputs, that cannot be used to fine-tune an algorithm's confidence. Second, ivis+GMM is a general approach and is not restricted to CXR reports. It is likely that application of this algorithm will yield comparable results in other free-text medical

records. Finally, ivis is a dimensionality reduction technique, which can be used to visualise complex data structures in two-dimensional space. It has been shown to scale linearly to millions of data points, resulting in more interpretable visualisations than comparable techniques such as t-distributed Stochastic Neighbour Embedding (t-SNE) [50].

Taken together, we have shown that supervised machine learning algorithms can reliably label free-text CXR radiological reports with excellent performance and using a relatively small number of training exemplars. More specifically, pre-training BiLSTM with fastText weights and the inclusion of the attention mechanism yields state-of-the-art accuracy that can be generalised to an independent testing set. To the best of our knowledge this is the first study where the generalisability of a machine learning algorithm for free-text CXR report interpretation has been demonstrated across two independently sourced and expert-labelled testing sets. Furthermore, we validate a general fully unsupervised approach that utilises Siamese Neural Networks and GMMs to reliably label large free-text corpora. Although direct application of expert knowledge to unlabelled radiograms remains the gold-standard of image annotation, we anticipate that our results can be used to effectively extract standardized clinical information from CXR radiological reports, facilitating large-scale training of modern clinical decision support systems for CXR triage.

Acknowledgments

We thank Dr. David Stobo for the invaluable clinical feedback throughout the study, Claire MacDonald for data extraction from NHS GGC SafeHaven, and James Blackwood for help with project coordination.

Author Contributions

Conceptualization: Ignat Drozdov, David J. Lowe.

Data curation: Daniel Forbes.

Formal analysis: Ignat Drozdov, Daniel Forbes, Benjamin Szubert, Mark Hall, Chris Carlin.

Funding acquisition: Ignat Drozdov, David J. Lowe.

Investigation: Ignat Drozdov, Mark Hall, Chris Carlin, David J. Lowe.

Methodology: Ignat Drozdov, Daniel Forbes, Benjamin Szubert, Mark Hall, Chris Carlin.

Project administration: Ignat Drozdov.

Resources: Ignat Drozdov, Mark Hall, David J. Lowe.

Software: Benjamin Szubert.

Supervision: Ignat Drozdov.

Validation: Ignat Drozdov, Benjamin Szubert, Mark Hall, Chris Carlin.

Visualization: Benjamin Szubert.

Writing – original draft: Daniel Forbes, Benjamin Szubert, Mark Hall, Chris Carlin, David J. Lowe.

Writing – review & editing: Daniel Forbes, Benjamin Szubert, Mark Hall, Chris Carlin, David J. Lowe.

References

1. Raoof S, Feigin D, Sung A, Raoof S, Irugulpati L, Rosenow EC 3rd. Interpretation of plain chest roentgenogram. Chest. 2012; 141(2):545–58. Epub 2012/02/09. https://doi.org/10.1378/chest.10-1302 PMID: 22315122.
2. Annarumma M, Withey SJ, Bakewell RJ, Pesce E, Goh V, Montana G. Automated Triaging of Adult Chest Radiographs with Deep Artificial Neural Networks. Radiology. 2019; 291(1):272. Epub 2019/03/22. https://doi.org/10.1148/radiol.2019194005 PMID: 30897046.
3. Rajpurkar P, Irvin J, Ball RL, Zhu K, Yang B, Mehta H, et al. Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. PLoS Med. 2018; 15(11):e1002686. Epub 2018/11/21. https://doi.org/10.1371/journal.pmed.1002686 PMID: 30457988 following competing interests: CPL holds shares in whiterabbit.ai and Nines.ai, is on the Advisory Board of Nuance Communications and on the Board of Directors for the Radiological Society of North America, and has other research support from Philips, GE Healthcare, and Philips Healthcare. MPL holds shares in and serves on the Advisory Board for Nines.ai. None of these organizations have a financial interest in the results of this study.
4. Irvin J, Rajpurkar P, Ko M, Yu Y, Ciurea-Ilcus S, Chute C, et al. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. arXiv e-prints [Internet]. 2019 January 01, 2019. https://ui.adsabs.harvard.edu/abs/2019arXiv190107031I.
5. Putha P, Tadepalli M, Reddy B, Raj T, Chiramal JA, Govil S, et al. Can Artificial Intelligence Reliably Report Chest X-Rays?: Radiologist Validation of an Algorithm trained on 2.3 Million X-Rays. arXiv e-prints [Internet]. 2018 July 01, 2018. https://ui.adsabs.harvard.edu/abs/2018arXiv180707455P.
6. Pasa F, Golkov V, Pfeiffer F, Cremers D, Pfeiffer D. Efficient Deep Network Architectures for Fast Chest X-Ray Tuberculosis Screening and Visualization. Sci Rep. 2019; 9(1):6268. Epub 2019/04/20. https://doi.org/10.1038/s41598-019-42557-4 PMID: 31000728
7. Rajpurkar P, Irvin J, Zhu K, Yang B, Mehta H, Duan T, et al. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. arXiv e-prints [Internet]. 2017 November 01, 2017. https://ui.adsabs.harvard.edu/abs/2017arXiv171105225R.
8. Ausawalaithong W, Marukatat S, Thirach A, Wilaiprasitporn T. Automatic Lung Cancer Prediction from Chest X-ray Images Using Deep Learning Approach. arXiv e-prints [Internet]. 2018 August 01, 2018. https://ui.adsabs.harvard.edu/abs/2018arXiv180810858A.
9. Sun C, Shrivastava A, Singh S, Gupta A. Revisiting Unreasonable Effectiveness of Data in Deep Learning Era. arXiv e-prints [Internet]. 2017 July 01, 2017. https://ui.adsabs.harvard.edu/abs/2017arXiv170702968S.
10. Vayena E, Blasimme A, Cohen IG. Machine learning in medicine: Addressing ethical challenges. PLoS Med. 2018; 15(11):e1002689. Epub 2018/11/07. https://doi.org/10.1371/journal.pmed.1002689 PMID: 30399149 following competing interests: EV has received speaking fees from SwissRe, Novartis R&D Academy, and Google Netherlands. IGC served as a consultant for Otsuka Pharmaceuticals advising on the use of digital medicine for its Abilify MyCite product. IGC is supported by the Collaborative Research Program for Biomedical Innovation Law, which is a scientifically independent collaborative research program supported by Novo Nordisk Foundation. AB served as a consultant for Celgene Corporation for the preparation of a workshop on pharmaceutical innovation and received honoraria from SwissRe for participating at an internal event on genome editing.
11. Oakden-Rayner L. Exploring large scale public medical image datasets. arXiv e-prints [Internet]. 2019 July 01, 2019. https://ui.adsabs.harvard.edu/abs/2019arXiv190712720O.
12. Pons E, Braun LM, Hunink MG, Kors JA. Natural Language Processing in Radiology: A Systematic Review. Radiology. 2016; 279(2):329–43. Epub 2016/04/19. https://doi.org/10.1148/radiol.16142770 PMID: 27089187.
13. Friedman C, Alderson PO, Austin JH, Cimino JJ, Johnson SB. A general natural-language text processor for clinical radiology. J Am Med Inform Assoc. 1994; 1(2):161–74. Epub 1994/03/01. https://doi.org/10.1136/jamia.1994.95236146 PMID: 7719797
14. Chapman BE, Lee S, Kang HP, Chapman WW. Document-level classification of CT pulmonary angiography reports based on an extension of the ConText algorithm. J Biomed Inform. 2011; 44(5):728–37. Epub 2011/04/05. https://doi.org/10.1016/j.jbi.2011.03.011 PMID: 21459155
15. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. J Biomed Inform. 2001; 34(5):301–10. Epub 2002/07/19. https://doi.org/10.1006/jbin.2001.1029 PMID: 12123149.
16. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. arXiv e-prints [Internet]. 2017 May 01, 2017. https://ui.adsabs.harvard.edu/abs/2017arXiv170502315W.

17. 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. Epub 2017/11/15. https://doi.org/10.1148/radiol.2017171115 PMID: 29135365.

18. Cornegruta S, Bakewell R, Withey S, Montana G. Modelling Radiological Language with Bidirectional Long Short-Term Memory Networks. arXiv e-prints [Internet]. 2016 September 01, 2016. https://ui.adsabs.harvard.edu/abs/2016arXiv160908409C.

19. Bustos A, Pertusa A, Salinas J-M, de la Iglesia-Vayá M. PadChest: A large chest x-ray image dataset with multi-label annotated reports. arXiv e-prints [Internet]. 2019 January 01, 2019. https://ui.adsabs.harvard.edu/abs/2019arXiv190107441B.

20. Lindsey R, Daluiski A, Chopra S, Lachapelle A, Mozer M, Sicular S, et al. Deep neural network improves fracture detection by clinicians. Proc Natl Acad Sci U S A. 2018; 115(45):11591–6. Epub 2018/10/24. https://doi.org/10.1073/pnas.1806905115 PMID: 30348771

21. Drew BJ, Harris P, Zegre-Hemsey JK, Mammone T, Schindler D, Salas-Boni R, et al. Insights into the problem of alarm fatigue with physiologic monitor devices: a comprehensive observational study of consecutive intensive care unit patients. PLoS One. 2014; 9(10):e110274. Epub 2014/10/23. https://doi.org/10.1371/journal.pone.0110274 PMID: 25338067

22. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. NPJ Digit Med. 2018; 1:18. Epub 2018/05/08. https://doi.org/10.1038/s41746-018-0029-1 PMID: 31304302

23. Johnson AEW, Pollard TJ, Berkowitz SJ, Greenbaum NR, Lungren MP, Deng C-y, et al. MIMIC-CXR: A large publicly available database of labeled chest radiographs. arXiv e-prints [Internet]. 2019 January 01, 2019. https://ui.adsabs.harvard.edu/abs/2019arXiv190107042J.

24. Szubert B, Cole JE, Monaco C, Drozdov I. Structure-preserving visualisation of high dimensional single-cell datasets. Sci Rep. 2019; 9(1):8914. Epub 2019/06/22. https://doi.org/10.1038/s41598-019-45301-0 PMID: 31222035

25. Hansell DM, Bankier AA, MacMahon H, McLoud TC, Muller NL, Remy J. Fleischner Society: glossary of terms for thoracic imaging. Radiology. 2008; 246(3):697–722. Epub 2008/01/16. https://doi.org/10.1148/radiol.2462070712 PMID: 18195376.

26. Doccano. Doccano: Open source text annotation tool for machine learning practitioner 2019 [cited 2019 August 1, 2019]. https://github.com/chakki-works/doccano.

27. Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching Word Vectors with Subword Information. arXiv e-prints [Internet]. 2016 July 01, 2016. https://ui.adsabs.harvard.edu/abs/2016arXiv160704606B.

28. Graves A, Fernández S, Schmidhuber J, editors. Bidirectional LSTM Networks for Improved Phoneme Classification and Recognition 2005; Berlin, Heidelberg: Springer Berlin Heidelberg.

29. Mullenbach J, Wiegreffe S, Duke J, Sun J, Eisenstein J. Explainable Prediction of Medical Codes from Clinical Text. arXiv e-prints [Internet]. 2018 February 01, 2018. https://ui.adsabs.harvard.edu/abs/2018arXiv180205695M.

30. Luong M-T, Pham H, Manning CD. Effective Approaches to Attention-based Neural Machine Translation. arXiv e-prints [Internet]. 2015 August 01, 2015. https://ui.adsabs.harvard.edu/abs/2015arXiv150804025L.

31. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention Is All You Need. arXiv e-prints [Internet]. 2017 June 01, 2017. https://ui.adsabs.harvard.edu/abs/2017arXiv170603762V.

32. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv e-prints [Internet]. 2018 October 01, 2018. https://ui.adsabs.harvard.edu/abs/2018arXiv181004805D.

33. Sanh V, Debut L, Chaumond J, Wolf T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv e-prints [Internet]. 2019 October 01, 2019. https://ui.adsabs.harvard.edu/abs/2019arXiv191001108S.

34. Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov R, Le QV. XLNet: Generalized Autoregressive Pretraining for Language Understanding. arXiv e-prints [Internet]. 2019 June 01, 2019. https://ui.adsabs.harvard.edu/abs/2019arXiv190608237Y.

35. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv e-prints [Internet]. 2019 July 01, 2019. https://ui.adsabs.harvard.edu/abs/2019arXiv190711692L.

36. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, et al. HuggingFace's Transformers: State-of-the-art Natural Language Processing. arXiv e-prints [Internet]. 2019 October 01, 2019. https://ui.adsabs.harvard.edu/abs/2019arXiv191003771W.

37. Joulin A, Grave E, Bojanowski P, Mikolov T. Bag of Tricks for Efficient Text Classification. arXiv e-prints [Internet]. 2016 July 01, 2016. https://ui.adsabs.harvard.edu/abs/2016arXiv160701759J.

38. Klambauer G, Unterthiner T, Mayr A, Hochreiter S. Self-Normalizing Neural Networks. arXiv e-prints [Internet]. 2017 June 01, 2017. https://ui.adsabs.harvard.edu/abs/2017arXiv170602515K.

39. Joachims T, editor Text categorization with Support Vector Machines: Learning with many relevant features 1998; Berlin, Heidelberg: Springer Berlin Heidelberg.

40. Kocbek S, Cavedon L, Martinez D, Bain C, Manus CM, Haffari G, et al. Text mining electronic hospital records to automatically classify admissions against disease: Measuring the impact of linking data sources. J Biomed Inform. 2016; 64:158–67. Epub 2016/10/16. https://doi.org/10.1016/j.jbi.2016.10.008 PMID: 27742349.

41. Weng WH, Wagholikar KB, McCray AT, Szolovits P, Chueh HC. Medical subdomain classification of clinical notes using a machine learning-based natural language processing approach. BMC Med Inform Decis Mak. 2017; 17(1):155. Epub 2017/12/02. https://doi.org/10.1186/s12911-017-0556-8 PMID: 29191207

42. Slack D, Friedler SA, Scheidegger C, Dutta Roy C. Assessing the Local Interpretability of Machine Learning Models. arXiv e-prints [Internet]. 2019 February 01, 2019. https://ui.adsabs.harvard.edu/abs/2019arXiv190203501S.

43. Tulio Ribeiro M, Singh S, Guestrin C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. arXiv e-prints [Internet]. 2016 February 01, 2016. https://ui.adsabs.harvard.edu/abs/2016arXiv160204938T.

44. Lai S, Xu L, Liu K, Zhao J. Recurrent convolutional neural networks for text classification. Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence; Austin, Texas. 2886636: AAAI Press; 2015. p. 2267–73.

45. Collobert R, Weston J, Bottou L, Karlen M, Kavukcuoglu K, Kuksa P. Natural Language Processing (almost) from Scratch. arXiv e-prints [Internet]. 2011 March 01, 2011. https://ui.adsabs.harvard.edu/abs/2011arXiv1103.0398C.

46. Thompson AJ. Re: The radiology report—are we getting the message across? Clin Radiol. 2012; 67 (7):723; author reply 4–5. Epub 2012/06/05. https://doi.org/10.1016/j.crad.2012.01.008 PMID: 22655596.

47. Mikolov T, Chen K, Corrado G, Dean J. Efficient Estimation of Word Representations in Vector Space. arXiv e-prints [Internet]. 2013 January 01, 2013. https://ui.adsabs.harvard.edu/abs/2013arXiv1301.3781M.

48. Guan Q, Huang Y, Zhong Z, Zheng Z, Zheng L, Yang Y. Diagnose like a Radiologist: Attention Guided Convolutional Neural Network for Thorax Disease Classification. arXiv e-prints [Internet]. 2018 January 01, 2018. https://ui.adsabs.harvard.edu/abs/2018arXiv180109927G.

49. Yang Z, Yang D, Dyer C, He X, Smola A, Hovy E, editors. Hierarchical Attention Networks for Document Classification 2016 jun; San Diego, California: Association for Computational Linguistics.

50. Maaten Lvd. Learning a Parametric Embedding by Preserving Local Structure. In: David van D, Max W, editors. Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics; Proceedings of Machine Learning Research: PMLR; 2009. p. 384–91.