Distinguishing between Dementia with Lewy Bodies (DLB) and Alzheimer's Disease (AD) using Mental Health Records: a Classification Approach

Zixu Wang[1], **Julia Ive**[2], **Sinéad Moylett**[3], **Christoph Mueller**[1,5], **Rudolf N. Cardinal**[3,4], **Sumithra Velupillai**[1], **John O'Brien**[3,4], and **Robert Stewart**[1,5]
[1]Institute of Psychiatry, Psychology & Neuroscience, King's College London, UK
[2]Department of Computing, Imperial College London, UK
[3]Department of Psychiatry, University of Cambridge, UK
[4]Cambridgeshire & Peterborough NHS Foundation Trust, UK
[5]South London and Maudsley NHS Foundation Trust, UK
{zixu.wang,christoph.mueller,sumithra.velupillai,robert.stewart}@kcl.ac.uk
j.ive@imperial.ac.uk, rnc1001@cam.ac.uk
{smm212,john.obrien}@medschl.cam.ac.uk

Abstract

While dementia with Lewy bodies (DLB) is the second most common type of neurodegenerative dementia following Alzheimer's disease (AD), it is difficult to distinguish from AD. We propose a method for DLB detection by using mental health record (MHR) documents from a (3-month) period before a patient has been diagnosed with DLB or AD. Our objective is to develop a model that could be clinically useful to differentiate between DLB and AD across various datasets from different healthcare institutions. We cast this as a classification task using convolutional neural network (CNN), an efficient neural model for text classification. We experiment with different representation models, and explore the features that contribute to model performances. In addition, we apply temperature scaling, a simple but efficient model calibration method, to produce more reliable predictions. We believe the proposed method has important potential for clinical applications using routine healthcare records, and for generalising to other relevant clinical record datasets. To the best of our knowledge, this is the first attempt to distinguish DLB from AD using mental health records, and to improve the reliability of DLB predictions.

1 Introduction

Alzheimer's disease (AD) is the most prevalent type of dementia, characterised by progressive cognitive impairment such as memory loss. Dementia with Lewy bodies (DLB), also known as Lewy body dementia, is the second most common type of neurodegenerative dementia following Alzheimer's disease (AD), with the defining features of fluctuating cognition, recurrent visual hallucinations, rapid eye movement (REM) sleep behaviour disorder, and Parkinsonian motor symptoms in addition to dementia (Walker et al., 2015). Particularly in the early stages, prior to diagnosis, DLB and AD are difficult to distinguish, hence the detection rates of DLB are sub-optimal, with a large proportion of cases missed or misdiagnosed as AD (Kane et al., 2018). Detection of DLB is, however, crucial as compared to AD and other forms of dementia (*e.g.* Parkinson's disease dementia (PDD)[1]). DLB has a worse prognosis across key outcomes such as mortality, hospitalisation, move into residential care, quality of life, and healthcare costs (Mueller et al., 2017). Moreover, not only is early diagnosis paramount, different types of treatments can have different impacts on these patient groups, *e.g.* antipsychotics, which adds to the importance of accurate and timely diagnoses.

Due to the challenges in recognising DLB clinically, it has been difficult to recruit large research cohorts of representative patients with DLB, and the increasing use of routinely collected healthcare data has been suggested as a potential solution to this shortage. Applying classical methods of symptom ascertainment using natural language processing (NLP) in routinely collected data is however difficult in patients with DLB, as clinicians tend to record the defining features only if they have also

[1]The distinction between DLB and PDD is largely around the degree of cognitive impairment and timing of motor symptoms, and they are on a continuum, hence the distinction is less clinically important in this case. Thus, we do not focus on this distinction here.

made the correct DLB diagnosis (Mueller et al., 2018). Therefore, we applied novel neural models of NLP to test whether these can be clinically useful to distinguish DLB and AD, and to provide assistance to mitigate expensive outcomes from misdiagnoses of DLB.

This task is challenging because DLB and AD share certain clinical and biological similarities that make them particularly difficult to differentiate. Motivated by the emergence of neural models and NLP methods applied to the biomedical domain, we cast this as a binary text classification task, where we use convolutional neural networks (CNNs) (LeCun et al., 1998; Krizhevsky et al., 2012; Kim, 2014) to address it. Additionally, the generalisation of well-trained models is notably more difficult, since different formats and grammatical patterns emerge in MHRs across different healthcare institutions. In order to test the efficiency of our proposed methodology, we use two datasets (CRIS[2] and CRATE[3]) from two different MHR (clinical documentation) systems and healthcare institutions, with the aim of comparing the model's performances on similar datasets containing relevant data, but with different contextual structures.

To assist the analysis of our experimental results, and to bridge the gap between model accuracy and confidence, we also study an approach where the model confidence estimates are calibrated. Confidence calibration is important for classification models. Classification networks must not only be accurate, but should also indicate when they are likely to be incorrect; a well-calibrated network matches its confidence to its accuracy so that it is confident when it is accurate, and uncertain when it is not. We use the calibration method named temperature scaling, where expected calibration error (ECE), the expectation of the differences between confidence and accuracy, is used as the primary empirical metric to measure calibration (Guo et al., 2017).

In this paper, we present our preliminary work towards automatically distinguishing individuals diagnosed with DLB or AD using neural network models and MHR texts. This methodology can provide an efficient technique for detecting and intervening DLB. Our contributions are threefold: 1) we introduce a CNN approach for the classification on DLB and AD using MHRs; 2) we investigate the performance of the proposed model on two MHR datasets from two different healthcare institutions with different formats and patterns; 3) we also apply a neural model calibration method to help in understanding when the model predictions tend to be brittle, so that the model can output confidence scores with higher reliability.

2 Related Work

With the success of neural models for many NLP tasks, deep learning methods, as well as word embeddings, have started to be applied to the biomedical and/or clinical domains (Cohen and Demner-Fushman, 2014; Wang et al., 2018; Kormilitzin et al., 2020) including mental health, such as automatic detection and classification of cognitive impairment.

For example, three neural models (CNNs-, LSTM-RNNs-, and CNN-LSTM-based) were applied to distinguish AD and Control patients from DementiaBank (Karlekar et al., 2018; Becker et al., 1994). CNN-LSTM model achieves state-of-the-art performance on the AD classification task. Since neural models are usually black-boxes and it is hard to interpret the reasoning for final classification decisions, various visualisation techniques have been proposed for neural networks (Mahendran and Vedaldi, 2015; Samek et al., 2016; Li et al., 2016; Kádár et al., 2017). Karlekar et al. (2018) illustrated two visualisation methods for interpretation, based on activation clustering and first-derivative saliency methods, to assist the analysis and consolidation of distinctive grammatical patterns of contextual information from AD patients.

Early detection plays a crucial part in the study of dementia. Pan et al. (2019) proposes a hierarchical model that encompasses both the hierarchical and sequential structures of picture description with attention mechanism, and detecting signs of cognitive decline at both the word and sentence levels, by using the DementiaBank and an in-house database of Cookie Theft picture descriptions (Mirheidari et al., 2017). Pan et al. (2019) shows both the proposed hierarchical structure and the attention

[2] *South London and Maudsley (SLaM) NHS Foundation Trust* Database - Clinical Record Interactive Search (CRIS) https://www.maudsleybrc.nihr.ac.uk/facilities/clinical-record-interactive-search-cris/

[3] *Cambridgeshire and Peterborough NHS Foundation Trust (CPFT)* Database - Clinical Records Anonymisation and Text Extraction (CRATE) http://www.psychiatry.cam.ac.uk/oap/research/lewy-cris/

mechanism contribute to the improvement in AD detection.

Most NLP studies addressing dementia use language transcripts from clinical cohorts, such as the DementiaBank (Becker et al., 1994). To our knowledge, very few studies have used MHR documents and NLP for modelling detection of dementia types, and we are not aware of any studies using NLP and MHRs for detection of DLB. McCoy Jr. et al. (2020) presents a study using electronic health record (EHR) data for stratifying risk for dementia onset, using a bespoke NLP approach for scoring symptoms in the clinical texts. This NLP approach, however, relies on pre-defined terms, and addresses a slightly different clinical problem.

When applying neural networks to real-world decision-making systems, classification networks must not only be accurate, but also should indicate when they are likely to be incorrect. A network should provide a calibrated confidence measure in addition to its prediction. Calibrated confidence estimates are also important for model interpretability. Guo et al. (2017) identify methods, which can alleviate miscalibrated problems in neural networks, and offer insight and intuition into network training and architectural trends that may cause miscalibration. Good confidence estimates can provide valuable extra information to establish trustworthiness in early detection of cognitive impairment.

3 Methodology

Our proposed approach uses a CNN model to distinguish DLB and AD patients. We compare the performance of using an embedding layer (`Emb-layer`) and pre-trained embeddings (`BioWord2Vec`) on our classification task, and finally apply a post-processing method (temperature scaling) for model calibration.

3.1 Input representation: word embeddings

We compare two approaches for the input, using high-dimensional word vectors (Mikolov et al., 2013): 1) a randomly initialised embedding layer and trained with the neural network, and 2) pre-trained biomedical word embeddings.

For the pre-trained embeddings, we use `BioWord2Vec`, distributed word representations proposed in Zhang et al. (2019)[4]. The biomedical word embeddings are learnt based on medical subject heading (MeSH) terms and text sequences, employing the fastText (Bojanowski et al., 2017) subword embedding model. `BioWord2Vec` outperforms the current state-of-the-art word embeddings in most BioNLP and/or ClinicalNLP tasks, suggesting that the sub-word information and domain knowledge are indeed able to improve the quality of biomedical word representations and better capture their semantics.

3.2 Convolutional Neural Network

We apply the convolutional neural network (CNN)[5] model (Kim, 2014) on our DLB and AD classification task. The input to the model are all documents of each patient concatenated and represented as a matrix using each of the embedding configurations. We use filters that slide over full rows of the matrix. The height of the filters may vary, but sliding windows over 3-5 words at a time are typical. Next, we max-pool (a sample-based discretisation process) the result of the convolutional layer into a long feature vector, add dropout regularisation, and the result is then passed to a softmax layer that outputs probabilities over two classes.

We use a logistic regression (`LR`) model as a baseline. Documents are pre-processed by tokenising and lowercasing. We compare two different text representations: bag-of-words (`BoW`) and term frequency-inverse document frequency (`TF-IDF`) counts. For `TF-IDF` counts, we selected a minimum document frequency of 5 and a maximum of 5,000 features.

3.3 Temperature Scaling

Temperature scaling is a post-processing technique which can almost perfectly restore network calibration (Guo et al., 2017), and can be easily added to any models. For classification problems, the neural network model outputs a vector known as the logits. The logits vector is passed through a softmax function to get class probabilities. Temperature scaling [6] simply divides the logits vector by a learnt scalar parameter, *i.e.*

$$P(\hat{\mathbf{y}}) = \frac{\exp(\mathbf{z}/T)}{\sum_j \exp(z_j/T)} \quad (1)$$

[4]During the preparation of this paper, more work on advanced pre-trained word embeddings emerged and we applied `BioWord2Vec`, one that was most relevant to our datasets.

[5]We first conducted the experiments using the BioBERT (Lee et al., 2019) classification model but it has a comparative worse performance due to (small) vocabulary overlap rates and larger clinical document lengths, as compared to the standard configurations in the BioBERT pre-training framework.

[6]https://github.com/gpleiss/temperature_scaling

where $\hat{\mathbf{y}}$ is the prediction, $\mathbf{z}$ is the logit, and T is the learnt parameter. T is learnt on the validation set, where T is chosen to minimise negative log-likelihood (NLL). Intuitively, temperature scaling simply softens the neural network outputs. This makes the network slightly less confident, which in turn makes the confidence scores reflect true probabilities.

This post-processing calibration method is applied on our DLB and AD classification task, to narrow the gap between model confidence and accuracy. The calibrated confidence provides further assistance when deciding whether the individual prediction might be reliable or incorrect.

A scalar summary statistic for calibration can be useful to compare two distributions: accuracy and confidence. The difference between accuracy and confidence is defined as:

$$\mathbb{E}_{\hat{P}}\left[\left|P(\hat{Y}=y|\hat{P}=p)-p\right|\right] \quad (2)$$

where $\hat{Y}$ is a class prediction, and $\hat{P}$ is its associated confidence, *i.e.* the probability of correctness.

In practice, the model predictions are grouped into M interval bins (each of size $\frac{1}{M}$). Expected calibration error (ECE) is computed as the weighted average of the bins' accuracy/confidence differences:

$$\mathbf{ECE}=\sum_{m=1}^{M}\frac{|B_m|}{N}\left|\text{acc}(B_m)-\text{conf}(B_m)\right| \quad (3)$$

where B_m is a set of indices where the prediction confidence of samples falls into the interval $\left(\frac{m-1}{M},\frac{m}{M}\right]$, and n is the total number of samples across all bins. Perfect calibration is achieved when $\mathbf{ECE}=0$, that is $\text{acc}(B_m)=\text{conf}(B_m)\ \forall$ bins m.

4 Materials and Experimental Setup

By applying two types of word embeddings (`Emb-layer` and `BioWord2Vec`) for word representations, convolutional neural network (CNN) for model training, and temperature scaling for model calibration, we investigated and evaluated the efficiency of our proposed methodology on three datasets (CRIS, CRIS†, and CRATE) from two healthcare institutions.

4.1 Datasets

We use mental health records (MHRs) from two different healthcare institutions (*SLaM* and *CPFT*), for which we have obtained ethical approval[7]. From each MHR database, we extract documents for patients diagnosed either with DLB or AD. CRIS and CRIS† origin from *SLaM*.

Dataset / # patients	DLB	AD
CRIS	90	750
CRIS†	90	90
CRATE	98	80

Table 1: Dataset Statistics - number of DLB and AD patients in each dataset. The ground truth for CRIS is extracted without human annotation.

Acquisition of ground truth differed for the two datasets. For CRIS, the MHRs are identified according to ICD-10 diagnoses whose titles contain the string "dementia". For CRATE from *CPFT*, an initial manual scan removed clear non-cases, after which a manual search was completed by two experienced clinicians, with knowledge of DLB diagnostic criteria and symptom presentation, in order to determine ground truth DLB cases. Cases were identified as ground truth DLB if a diagnosis had been given by a clinician within the healthcare institution and was the most recent recorded diagnosis within the MHR (see (Price et al., 2017) for more details on data collection for CRATE). To have a more comparable dataset from *SLaM*, we also created CRIS†, in which we randomly selected AD cases from CRIS to obtain a more balanced distribution, while the DLB cases remain identical to CRIS.

Within each dataset, we have information about the `Patient_ID` and the `Diagnosis_Date` of DLB and AD patients respectively. For each patient with any of these diagnoses, we use only the text written *upon the first consultation until the date 3 months before the diagnosis* (concatenated into one document). The intuition is that we would like to remove MHRs closer to the date of diagnosis that could be more informative of the two diseases, and hence making the differentiation using NLP trivial. There is a total of 90 DLB patients and 750 AD patients in CRIS[8], and 98 DLB patients and 80 AD

[7]Data were obtained from the *SLaM* research database CRIS; and the *CPFT* Research Database (NHS Research Ethics 17/EE/0442), containing de-identified clinical data from electronic mental health records of Cambridgeshire Peterborough NHS Foundation Trust (*CPFT*).

[8]The distribution of DLB and AD patients from CRIS is close to the real distribution because diagnosed DLB is currently about 5% of all dementias and there is evidence that DLB should be around 10%, AD is around 70% (Mueller et al.,

Datasets	Max	Min	Median
CRIS	206,228	319	4,406
CRIS†	187,438	463	4,243
CRATE	733,388	28	2,710

Table 2: Statistics of document length, where **Max**, **Min**, and **Median** refer to the number of words of the document.

Datasets	Vocabulary size	Overlap
CRIS	186,002	47,360 (25.5%)
CRIS†	65,444	31,343 (47.9%)
CRATE	66,785	20,220 (29.9%)

Table 3: Statistics of vocabulary (`BioWord2Vec` contains 2,324,849 distinct words in total where 2,309,172 words come from the PubMed and 15,677 from MeSH.). The reason that there is a larger overlap in CRIS† might be from higher contextual consistency between CRIS† and `BioWord2Vec`.

patients in CRATE[9] (see Table 1). In CRIS†, the AD cases were extracted randomly from CRIS with the aim of making the results more comparable by equalising the number of DLB and AD patients (closer to the distribution in CRATE).

The length of each document varies in the datasets, ranging from tens of words to hundreds of thousands (see Table 2). On average, documents are longer in CRIS and CRIS†. Since the standard CNN model used for text classification takes the maximum length of samples as the uniform length, we considered to use only the median length of documents for more reasonable document lengths and save the computing capacities [10] (as shown in Table 2) to pad/cut documents to the same length, and use the latest diagnosis records as the training samples if the document exceeds the median.

4.2 Experimental Setup

In our binary classification task we consider DLB cases as positive and AD cases as negative. We preprocess the datasets by lowercasing and tokenising using regular expression operations. We use 5-fold cross-validation (CV) to segment the training datasets and ensure that particular subgroups have no deterministic effect on final model performance. All our models use an Adam optimizer (Kingma and Ba, 2014), with a learning rate of 0.001. We used a 2-D CNN. Filter sizes of $[3, 4, 5]$ were used with 128 filters per filter size. Batch size was set to 32. To avoid overfitting, we apply dropout to the output of all the functional layers (Srivastava et al., 2014), with the dropout rate set to 0.5. The final criteria are calculated by averaging the 5-fold cross-validation results.

In the ablation study, we remove important words from the training data and to trace changes in model performance. These important words are either the most informative of DLB and AD (*e.g.* Model B where a list of terms, expressions, and abbreviations related to the diagnoses of DLB and AD; and was composed manually), or obtained from our baseline model which contribute the most to the `LR` predictions (Model C). We believe these words are also indicative to neural models. Four models are designed and compared:

- **Model A:** The training data are the raw text for all the datasets.
- **Model B:** "lewy", "body", "bodies", "dlb", "ad", "lbd", "dementia" are removed from original text.
- **Model C:** "parkinson", "hallucinations", "visual", "symptoms" are removed from original text.
- **Model D:** Words mentioned in **Model B** and **Model C** are all removed from original text.

We use the temperature scaling calibration method, which does not affect the model's accuracy. We would want the confidence estimates (output probabilities) to be calibrated. For example, given 100 predictions, each with confidence of 0.8, we expect that 80 should be correctly classified. A perfect calibration should be an identity function between accuracy and confidence. We decide to measure calibration by using expected calibration error (ECE).

4.3 Evaluation

In order to test the efficiency of our model, we report the performances based on precision, recall, and F1-score. All the reported results are the average of 5-fold cross-validation (CV). We also report F1-scores for each fold. In addition, to better understand the underlying data, we extract the top-20

2017).

[9]The more balanced distribution of CRATE is an outcome of the manual extraction.

[10]For the CNN model, we use the sequence length 4,406 for CRIS and 2,710 for CRATE; for CRIS†, we applied the same median length (2,710) as CRATE, in order to make the results more comparable.

Datasets	Model	Word Representation	Precision	Recall	F1-score
CRIS	LR	BoW	0.76	0.63	0.66 (0.48, 0.72, 0.73, 0.67, 0.70)
		TF-IDF	0.91	0.52	0.49 (0.51, 0.44, 0.50, 0.67, 0.34)
	CNN	Emb-layer	**0.92**	**0.85**	**0.87 (0.91, 0.87, 0.82, 0.87, 0.88)**
		BioWord2Vec	0.75	0.55	0.63 (0.11, 0.78, 0.58, 0.91, 0.80)
CRIS†	LR	BoW	0.71	0.67	0.68 (0.69, 0.72, 0.80, 0.54, 0.65)
		TF-IDF	0.75	0.75	0.75 (0.72, 0.75, 0.72, 0.80, 0.75)
	CNN	Emb-layer	**0.87**	0.81	0.73 (0.63, 0.78, 0.76, 0.77, 0.74)
		BioWord2Vec	0.76	**0.85**	**0.78 (0.81, 0.68, 0.83, 0.80, 0.82)**
CRATE	LR	BoW	0.75	0.65	0.69 (0.78, 0.75, 0.52, 0.59, 0.81)
		TF-IDF	0.71	**0.85**	**0.77 (0.75, 0.69, 0.79, 0.84, 0.78)**
	CNN	Emb-layer	**0.88**	0.59	0.70 (0.68, 0.67, 0.71, 0.73, 0.73)
		BioWord2Vec	0.63	0.82	0.71 (0.71, 0.71, 0.71, 0.69, 0.73)

Table 4: DLB classification results (CRIS, CRIS†, and CRATE), using a logistic regression (LR) model with bag-of-words (BoW) or TF-IDF counts representation, and using CNN with embeddings from the training data (Emb-layer) or pre-trained embeddings (BioWord2Vec): precision, recall, and F1-score, average from 5-fold cross-validation (F1-scores for each fold are shown in brackets).

Dataset	LR	F1-score	Features (words)
CRIS	BoW	0.66	**hallucinations**, [person name], today, night, [person name], currently, [person name], body, **symptoms**, [person name], **parkinson**, score, continues, review
	TF-IDF	0.49	**hallucinations**, rivastigmine, **parkinson**, formcheckbox, lithium, quetiapine, [person name], reg, **visual**, [person name], [person name], night
CRATE	BoW	0.69	[person name], **hallucinations**, mr, place, change, opmh, allowance, [person name], [person name], note, stanground, [person name], time, [person name], mental
	TF-IDF	0.77	mr, **hallucinations**, **parkinson**, [person name], care, ext, liaison, transfer, mood, able, risk, review, **visual**, carers, **symptoms**, admission, lodge, [person name]

Table 5: Top-20 words contributing the most to the DLB detection using logistic regression (LR) with BoW and TF-IDF counts representation (a minimum document frequency of 5 and a maximum of 5,000 features).

words contributing the most to the DLB classification in the LR model with both the BoW and TF-IDF counts representations.

5 Results

Overall classification results are reported in Table 4. Two kinds of word representations are used with the LR model: BoW and TF-IDF. Using BoW features resulted in higher F1-score (0.66) as compared to TF-IDF features (0.49) for CRIS; while the opposite is observed for CRATE (0.69 for BoW and 0.77 with TF-IDF features). In general, CNN achieves better results compared to the baseline LR (0.87 for CNN with Emb-layer on CRIS), and lower deviation for each fold in 5-fold cross-validation. On CRATE, the LR model with TF-IDF features performs best (0.77).

Comparing the performances of random initialised word embeddings (Emb-layer) and pre-trained BioWord2Vec, the result using Emb-layer achieves higher F1-score (0.87) than BioWord2Vec (0.63) for CRIS. Results on CRATE using Emb-layer and BioWord2Vec are, on the other hand, quite close considering F1-scores and their stabilities for 5-fold CV.

However, for CRIS†, using pre-trained embeddings BioWord2Vec (0.78) performs better than Emb-layer (0.73), with more comparable data sizes of DLB and AD. Our proposed model CNN with BioWord2Vec achieves the highest F1-score (0.78) among four models (LR with BoW and TF-IDF, CNN with Emb-layer and BioWord2Vec). With the same settings, the

Datasets	CNN	A	B	C	D
CRIS	Emb-layer	0.879	0.490	0.810	0.321
	BioWord2Vec	0.636	0.633	0.543	0.500
CRIS†	Emb-layer	0.738	0.652	0.730	0.733
	BioWord2Vec	0.784	0.785	0.637	0.700
CRATE	Emb-layer	0.703	0.649	0.692	0.667
	BioWord2Vec	0.712	0.702	0.690	0.640

Table 6: Comparison of models (F1-scores) with different input texts.

F1-score is also higher than that of CRIS (0.63) with lower deviation, which might be the outcome of a more balanced dataset. In comparison to CRATE (0.71), although the F1-score is slightly higher, the CRATE has a comparatively lower deviation. This result might be inherited from the fact that there is a significant increase in the overlap between CRIS† and CRATE datasets (see Table 3).

We also report the top-20 most important features contributing to the prediction in the LR model using BoW and TF-IDF counts representations (see Table 5). It is obvious that "hallucinations", "parkinson", "visual", "symptoms" are ranked highly in both CRIS and CRATE.

Inspired by the important features from LR, our baseline method, we removed the top-ranked important words from the pilot training data. We observed that after removing the core dementia-related words we still obtain similar F1-scores for CRATE using both types of embeddings (see Table 6: models B-D compared to A). These words, however, seem to contribute more to the predictions of CRIS patients and as informative as DLB symptoms in this case. Results for CRIS† indicate the efficiency of a more balanced dataset and higher vocabulary overlap with BioWord2Vec, where we obtained less performance decrease when removing informative words. This would imply the remainder sets of words could also contribute to the model predictions.

Using **Model A** as the base model for model calibration, where raw text serves input to our CNN model with the BioWord2Vec word representation, we obtain well-calibrated model for all CRIS, CRIS†, and CRATE (see Table 7).

It is worth noting that models trained on three datasets experience some degrees of miscalibration. (1) The confidences of two models (before and after calibration) decrease from over-confident to a reliable level after temperature scaling. The difference between two confidence scores indicates the performance of calibration and the model's stability. If the confidence level drops significantly (for instance, CRATE), this means there is more uncertainty in the calibrated model estimates, but less gap between accuracy and confidence. (2) According to Guo et al. (2017), the ECE is typically between 4% to 10% on benchmark datasets. In our experiment, we expected the scores of ECE to be higher, as MHRs are much more free-formed and noisy. Through the comparison of ECE before and after calibration, we can observe that temperature scaling does calibrate on the datasets, which is also supported by the reduction in confidence and NLL. (3) The NLL is often used to define how well a neural network classifies data. A high NLL means the classification is inaccurate. A low NLL otherwise indicates the prediction matches that of the expected value. The NLL decrease in our models on the datasets means that the calibration produces more reliable prediction outputs.

6 Discussion

To our knowledge, this is the first study on automatically distinguishing dementia with Lewy bodies (DLB) from Alzheimer's disease (AD) using MHRs. We investigated the performance of CNN models using different embedding representations on MHRs from two different healthcare institutions, and incorporating the method of model calibration into DLB classification to obtain reliable predictions.

To be able to apply NLP models to real-world biomedical tasks, we need first to embrace the challenges of the datasets. In our case, we face a range of such challenges: small data size, hence reduced reliability of predictions; class imbalance; noisy data; and contextual differences between datasets. These might be the reasoning behind higher deviation and instability of F1-scores observed in some predictions (see Table 4).

We attempted to mitigate these challenges by using a set of fairly standard techniques. We use 5-fold CV to ensure that every example appears during both training and testing. Using 5-fold CV, important information is more likely to be learnt, and consequently obtaining better approximations and enhancing robustness, whereas with larger datasets there is more chance to have a proper distribution of information for both training and testing.

Since most MHRs are written with different formats and grammatical patterns, we considered using pre-trained biomedical word embeddings

Datasets	Confidence	ECE (%)	NLL
CRIS	$0.97 \rightarrow 0.87$ ($\Delta = 0.10$)	$8.234 \rightarrow 4.701$	$7.335 \rightarrow 3.027$
CRIS†	$0.88 \rightarrow 0.81$ ($\Delta = 0.07$)	$11.627 \rightarrow 9.525$	$2.853 \rightarrow 1.531$
CRATE	$0.88 \rightarrow 0.65$ ($\Delta = 0.23$)	$15.745 \rightarrow 8.351$	$1.089 \rightarrow 0.633$

Table 7: Model Calibration on CNN/BioWord2Vec combination (Before Calibration $\rightarrow$ After Calibration).

(`BioWord2Vec`) to get a unified word representation across different datasets. Those embeddings helped our models to rely less on explicit indicators of diagnoses (*e.g.* direct mentions of a diagnosis) while producing predictions and stabilised performance over cross-validation splits. However, using those embeddings might be hindered by excessive noise (concatenation of words and punctuation, misspellings) in data and hence poor vocabulary overlap. Better performance in this case can naturally be achieved if more in-domain data is available and embeddings are trained from scratch.

Finally, to improve the reliability of model predictions, temperature scaling, a simple but efficient calibration method, is used to narrow the gap between accuracy and confidence. The ECE scores from both before and after calibrations are used as the primary measures of model calibration. The well-calibrated model decreases in confidence. This can reflect the true probability of model predictions, and can provide a good assistance and reference when evaluating the model outputs.

Our proposed model and calibration method could prove useful clinically. Currently in clinical care there is a high level of under-diagnosis as well as lack of confidence in making a DLB diagnosis. Moreover, appropriate treatment is crucial, *e.g.* it is important to avoid antipsychotic prescribing for this patient group. Although the F1-scores and calibration results are not always perfect, they indicate that using routine healthcare data could be valuable for predictive model development even in cases where it is hard to obtain large datasets.

7 Conclusion and Future Work

In this paper, we propose to use a CNN approach for the task of detecting DLB patients by distinguishing them from AD patients. Our well-calibrated models are relatively robust after using temperature scaling, where calibrated probabilities are more informative of good probability estimates and true predictions. The proposed model is investigated on two MHR datasets from two different healthcare institutions, and achieves competitive results using two types of embeddings (`Emb-layer` and `BioWord2Vec`). The pre-trained biomedical word embeddings (`BioWord2Vec`) are efficient for all three datasets whilst CRIS relies much more on in-domain word distributions. In particular, `BioWord2Vec` can achieve lower deviations on model performance in ablation study.

Future work will be focused on the effectiveness of contextualised embeddings for a more general methodology where the detection of DLB can be realised across healthcare institutions. We would also like to investigate more effective pre-processing techniques to purify and clean the raw texts before feeding into advanced models, and to mitigate the noise commonly existed in health records. The code is available at `https://github.com/zixuwang1996/dlb_ad_classification`.

Acknowledgments

The work was funded by an Alzheimer's Society Biomedical Grant and the UK National Institute of Health Research (NIHR) Cambridge Biomedical Research Centre. RS is part-funded by: i) the National Institute for Health Research (NIHR) Biomedical Research Centre at the South London and Maudsley NHS Foundation Trust and King's College London; ii) a Medical Research Council (MRC) Mental Health Data Pathfinder Award to King's College London; iii) an NIHR Senior Investigator Award; iv) the National Institute for Health Research (NIHR) Applied Research Collaboration South London (NIHR ARC South London) at King's College Hospital NHS Foundation Trust. RNC's research is supported by the UK Medical Research Council (grant MC_PC_17213 to RNC). The *CPFT* Research Database is supported by the UK National Institute of Health Research (NIHR) Cambridge Biomedical Research Centre. We thank Cambridgeshire and Peterborough NHS Foundation Trust (*CPFT*) for supports in using the Clinical Records Anonymisation and Text Extraction (CRATE) system.

We would like to thank all anonymous reviewers for their helpful comments. The views expressed

are those of the author(s) and not necessarily those of the NHS, the NIHR, nor the Department of Health and Social Care.

References

James T. Becker, François Boiler, Oscar L. Lopez, Judith Saxton, and Karen L. McGonigle. 1994. The Natural History of Alzheimer's Disease: Description of Study Cohort and Accuracy of Diagnosis. *Archives of Neurology*, 51(6):585–594.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Kevin Bretonnel Cohen and Dina Demner-Fushman. 2014. *Biomedical natural language processing*, volume 11. John Benjamins Publishing Company.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330, International Convention Centre, Sydney, Australia. PMLR.

Ákos Kádár, Grzegorz Chrupała, and Afra Alishahi. 2017. Representation of linguistic form and function in recurrent neural networks. *Computational Linguistics*, 43(4):761–780.

Joseph PM Kane, Ajenthan Surendranathan, Allison Bentley, Sally AH Barker, John-Paul Taylor, Alan J Thomas, Louise M Allan, Richard J McNally, Peter W James, Ian G McKeith, et al. 2018. Clinical prevalence of lewy body dementia. *Alzheimer's research & therapy*, 10(1):19.

Sweta Karlekar, Tong Niu, and Mohit Bansal. 2018. Detecting linguistic characteristics of Alzheimer's dementia by interpreting neural models. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 701–707, New Orleans, Louisiana. Association for Computational Linguistics.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. Cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015.

Andrey Kormilitzin, Nemanja Vaci, Qiang Liu, and Alejo Nevado-Holgado. 2020. Med7: a transferable clinical natural language processing model for electronic health records. *arXiv preprint arXiv:2003.01271*.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*.

Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. 2016. Visualizing and understanding neural models in NLP. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 681–691, San Diego, California. Association for Computational Linguistics.

Aravindh Mahendran and Andrea Vedaldi. 2015. Understanding deep image representations by inverting them. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Thomas H. McCoy Jr., Larry Han, Amelia M. Pellegrini, Rudolph E. Tanzi, Sabina Berretta, and Roy H. Perlis. 2020. Stratifying risk for dementia onset using large-scale electronic health record data: A retrospective cohort study. *Alzheimer's & Dementia*, n/a(n/a). Publisher: John Wiley & Sons, Ltd.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

Bahman Mirheidari, DJ Blackburn, Kirsty Harkness, Traci Walker, Annalena Venneri, Markus Reuber, and Heidi Christensen. 2017. An avatar-based system for identifying individuals likely to develop dementia. In *Interspeech 2017*, pages 3147–3151. ISCA.

Christoph Mueller, Clive Ballard, Anne Corbett, and Dag Aarsland. 2017. The prognosis of dementia with lewy bodies. *The Lancet Neurology*, 16(5):390–398.

Christoph Mueller, Gayan Perera, Anto P Rajkumar, Manorama Bhattarai, Annabel Price, John T O'Brien, Clive Ballard, Robert Stewart, and Dag Aarsland. 2018. Hospitalization in people with dementia with lewy bodies: Frequency, duration, and cost implications. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 10:143–152.

Yilin Pan, Bahman Mirheidari, Markus Reuber, Annalena Venneri, Daniel Blackburn, and Heidi Christensen. 2019. Automatic hierarchical attention neural network for detecting ad.

Annabel Price, Redwan Farooq, Jin-Min Yuan, Vandana B Menon, Rudolf N Cardinal, and John T O'Brien. 2017. Mortality in dementia with lewy bodies compared with alzheimer's dementia: a retrospective naturalistic cohort study. *BMJ Open*, 7(11).

Wojciech Samek, Alexander Binder, Gregoire Montavon, Sebastian Lapuschkin, and Klaus Muller. 2016. Evaluating the visualization of what a deep neural network has learned. *IEEE Transactions on Neural Networks and Learning Systems*.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958.

Zuzana Walker, Katherine L Possin, Bradley F Boeve, and Dag Aarsland. 2015. Lewy body dementias. *The Lancet*, 386(10004):1683–1697.

Yanshan Wang, Sijia Liu, Naveed Afzal, Majid Rastegar-Mojarad, Liwei Wang, Feichen Shen, Paul Kingsbury, and Hongfang Liu. 2018. A comparison of word embeddings for the biomedical natural language processing. *Journal of biomedical informatics*, 87:12–20.

Yijia Zhang, Qingyu Chen, Zhihao Yang, Hongfei Lin, and Zhiyong Lu. 2019. Biowordvec, improving biomedical word embeddings with subword information and mesh. *Scientific data*, 6(1):1–9.