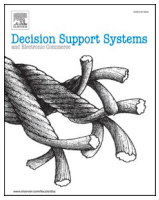
Improving triaging from primary care into secondary care using heterogeneous data-driven hybrid machine learning

Decision Support Systems 166 (2023) 113899

Contents lists available at ScienceDirect

Decision Support Systems

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



Improving triaging from primary care into secondary care using heterogeneous data-driven hybrid machine learning

Bing Wang [a], Weizi Li [a,*], Anthony Bradlow [b], Eghosa Bazuaye [c], Antoni T.Y. Chan [b]

[a] Informatics Research Centre, Henley Business School, University of Reading, Reading RG6 6UD, UK
[b] Rheumatology Department, Royal Berkshire NHS Foundation Trust, Reading RG1 5AN, UK
[c] Informatics Department, Royal Berkshire NHS Foundation Trust, Reading RG1 5AN, UK

ARTICLE INFO

Keywords:
Machine learning
Primary to secondary care triage
Ensemble method for heterogeneous data
Prediction explanation
NLP

ABSTRACT

Effective and rapid triaging from primary care into secondary care plays a pivotal role in providing patients with timely treatment and managing increasing demands for healthcare resources. Existing triaging methods from primary care to secondary care are labor-intensive processes that involve manually reviewing referral data from multiple sources and can cause long referral to treatment time. There has been no research using machine learning methods that automatically analyzes heterogeneous data including referral letters to recognize regularities to support the primary to secondary care triage. In this paper, we propose a heterogeneous data-driven hybrid machine learning model including Natural Language Processing (NLP) to improve hospital triage efficiency at the point of triage. The proposed model achieved a precision of 0.83, recall of 0.82, F1-Score of 0.83, accuracy of 0.82, AUC of 0.90 in identifying patients with non-inflammatory conditions (NIC) and inflammatory arthritis (IA) at the point of triage with explainable risk stratifications. Our model is piloted in a real-world trial in a large secondary care hospital in the UK to compare referral accuracy and time saved between our model and clinicians, and evaluate its acceptability by users. Our model achieved precision and recall of 0.83 and 0.81, compared with the precision and recall of 0.80 and 0.78 by clinicians. The research also shows that our model enabled decision support can save clinicians 8 h per week in assessing the referral assessment. This paper is the first study to streamline hospital triage from primary care to secondary care using machine learning.

1. Introduction

Rapid triage and referral assessment from primary to secondary care is essential for timely medical intervention to prevent death and disability [1]. However current manual referral assessment at secondary care hospitals is a time-consuming process and referrals from GP for some diseases are mostly inaccurate due to vague presenting symptoms. For example, early inflammatory arthritis (EIA) can be difficult to diagnose and can present with non-specific symptoms [2]. In the UK, many patients referred by General Practice (GP) ultimately did not have a diagnosis of EIA, with only 40% of referrals proved to be EIA cases entering EIA pathways at secondary hospitals in the period of 2019–2020 [3]. Inaccurate referrals can lead to longer times for elective care patients to access the right clinics. To improve the referral triage quality, every GP referral letter and clinical information need to be read and assessed by a specialist clinician in the secondary care hospital to determine the appropriate care pathway as shown in Fig. 1. This is a compulsory requirement for secondary hospitals in the UK before booking any appointment with patients in elective care [4] because evidence shows incorporating clinicians' assessments and feedback from secondary care hospitals will improve referral quality [5]. However, referral assessment is time-consuming for clinicians and it has to be fitted in the time that could have been better used for patient care and other clinical activities at the hospital. There have been delays between a hospital rheumatology department receiving a referral from GP for suspected EIA and the date of clinic assessment. The National Early Inflammatory Arthritis Audit (NEIAA) requires referral triage assessment by secondary care rheumatology department within three weeks of referral but fewer than half of hospitals achieve target times [1].

To improve the triage referral assessment, there are solutions such as referral guidelines, education interventions (e.g., feedback of referrals from secondary care clinicians to GP) and organizational interventions (e.g., the establishment of the referral management center), and financial incentives [5]. Those approaches are effective in improving referrals

* Corresponding author.
E-mail address: weizi.li@henley.ac.uk (W. Li).

https://doi.org/10.1016/j.dss.2022.113899
Received 29 March 2022; Received in revised form 7 November 2022; Accepted 9 November 2022
Available online 14 November 2022

to some extent however the cost-effectiveness remains unclear and can lead to higher costs according to the evidence [5,6]. Streamlining the referral assessment using machine learning becomes a key attention point given its potential to provide rapid assessment and decision support automatically without significant organizational cost. Machine learning-based decision support systems that identify patients' diseases based on healthcare data at the point of triaging have proven cost-effective and showed high accuracy in disease early diagnosis [7,8]. Therefore, our motivation is to utilize machine learning to provide a more cost-effective, faster, and more accurate referral assessment method for elective care. However real-world healthcare data at the point of triaging in elective care are sometimes incomplete and always heterogeneous [9] with different modalities, including unstructured data that is GP referral letters of clinical information summary and structured blood test results data. Extracting information from unstructured GP letters requires extensive feature engineering but state-of-the-art natural language processing (NLP) methods have the potential to automatically extract context information. To our best knowledge, there is no prior research using unstructured data contained in the GP letters for referral triage assessment, though GP referral letters often contain useful information and would be very helpful to be incorporated into machine learning models for triage efficiency improvement. Furthermore, the research of utilizing deep learning and natural language processing to process real-world heterogeneous data for referral triage from primary to secondary care is still a blank research field. In this research, we develop a novel heterogeneous data-driven hybrid machine learning approach to improve triaging from primary care to secondary care with the further contributions:

- We develop a hybrid machine learning method to address heterogeneity and incompleteness challenges in a real-world referral triaging context. We firstly integrated GP referral letters into modeling and developed an ensemble decision-making methodology for hospital referral triage. Specifically, our approach can cover all referral data scenarios by developing two models separately for patients either having GP referral letters or blood test data, and a probabilistic fusion method for patients having both heterogeneous data.
- We contribute to analytical methods of GP referral letters by developing a Bidirectional Encoder Representations from Transformers (BERT)-based dynamical feature fusion model to identify patients with inflammatory arthritis and patients with non-inflammatory conditions using GP referral letters.
- We develop local prediction explanation method into our triaging model to provide explainable referral triage recommendations, which will help clinicians to understand the underlying logic and ensure that the model can be checked for the reliability of model recommendations. This will also speed up the triaging process by highlighting the key information for clinicians.

2. Related works

We review the state-of-the-art hospital triage methods through the lens of traditional expert-based methods and data-driven methods respectively. A spectrum of methods from manual scoring methods to machine learning-based methods have been examined to identify research gaps in improving the referral triaging process and the contributions of our research in addressing current gaps.

2.1. Expert-based hospital triage research

Expert-based triage has a long history [10] and still plays a critical role in the hospital to accurately prioritize the patients' health care demands or triage patients when they arrive at the hospital [11,12]. This clinical assessment process is often conducted manually by hospital workers such as nursing staff [11], and sometimes performed online or remotely [13,14].

A fundamental step of this traditional triage method is to build the assessment scales based on the experts' experience. The majority of current expert-based triage methods are developed and used in the emergency department to identify urgency levels, such as Ipswich Triage Scale (ITS) [15], Australasian Triage Scale (ATS) [16], Manchester Triage Scale (MTS) [17], Canadian Triage and Acuity Scale (CTAS) [18], Emergency Severity Index (ESI) [19]. Furthermore, some early warning tools are used in the emergency department to identify patient deterioration in emergency care, such as Early Warning Scores (EWS) [20], Modified Early Warning Score (MEWS) [21], and National Early Warning Score (NEWS) [22]. However, the use of the early warning system such as the MEWS score has been controversial for its ability to escalate patients and cannot be used as the only source for ED triage referral in practice [23].

Existing triage scores designed for emergency care are not fit for purpose of the elective care which is our research focus. Elective care is non-urgent care normally referred by GP to different hospital specialist departments according to patients presenting symptoms and test results.

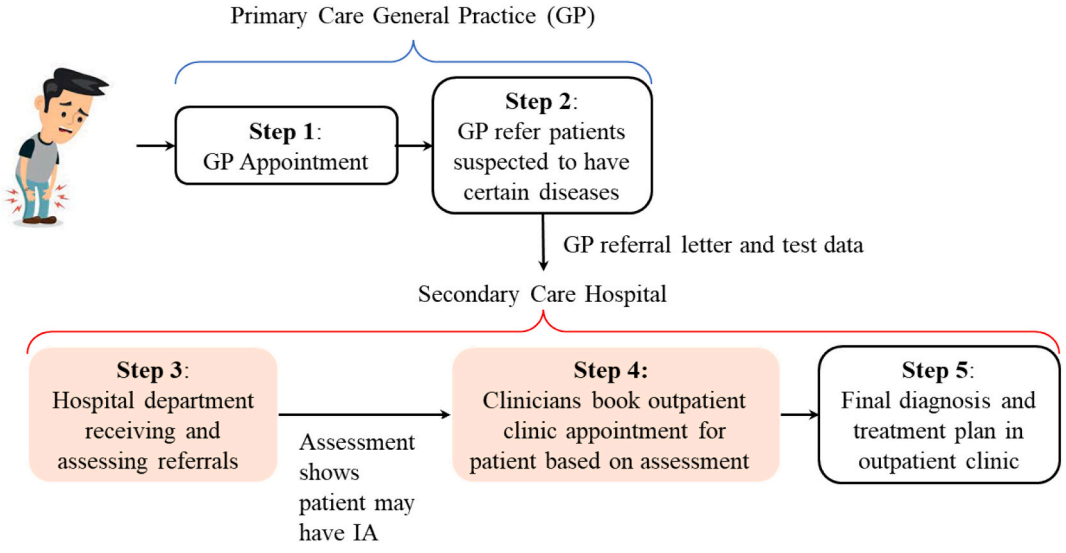


Fig. 1. Elective care referral triage process from primary care to secondary care hospital.

The referral of elective care from primary care GP to secondary care hospital is quite different from emergency care because the referral decisions rely heavily on disease-specific symptoms and biomarkers instead of urgency scales or deterioration levels [24,25]. There are some special triage scales developed for specific diseases. For example, assessment scales for obstetrics include United States Maternal Fetal Triage Index (MFTI) [26], Canada Obstetric Triage Acuity Scale (OTAS) [27], United Kingdom Birmingham Symptom specific Obstetric Triage System (BSOTS) [28], Iranian Obstetric Triage Index (IOTI) [29], Swiss Emergency Triage Scale (SETS) [30]. However, a manual referral assessment review is still needed to ensure triage accuracy, especially for disease like inflammatory arthritis where symptoms are non-specific and there is no single current marker that is diagnostically definitive [31–33]. Besides, tele-triage [13,34] is increasingly attracting attention because it to some extent can reduce the need to travel to clinics [11,35–37]. However, both hospital triage and tele-triage involve a heavy workload with manual reviewing and sometimes inaccurate triage to treatment. Manual assessment is inevitable in elective care because key clinical information is in the unstructured GP referral letters and currently there is no research to automatically risk stratify patient from unstructured letter.

2.2. Data-driven hospital triage research

With the development of artificial intelligence, data-driven methods show great potential in the applications of hospital triage. Therefore, there is an obvious shift from expert-driven to data-driven methods using machine learning and deep learning.

A primary application of machine learning is the prediction of mortality and hospitalization admission at the point of triage in the emergency department. Rocío Sanchez-Salmeron summarized machine learning-related applications in the Emergency Department (ED) triage [11]. Specifically, this study depicts that the machine learning models could outperform the traditional methods in different ED triage scenarios, including the 24-h mortality prediction [38], the early and short-term mortality prediction [39], and the critical care and hospitalization admission predictions [40–43]. Similarly, Logistic Regression (LR) has been used to predict the mortality of inpatients using information collected from the ED [44] as well as the inpatient admission from the emergency department [45]. Besides, Arnaud et al. used deep learning to predict hospitalization at the ED [46], and Tahayori et al. utilized machine learning married with the NLP method to predict the disposition of patients and thus optimize the resource allocation in the ED on the basis of emergency triage notes [47].

Machine learning has also been applied to the classification of the severity of the patients and early recognition of some specific diseases at the initial ED triage. For example, Zmiri et al. applied Naïve Bayes and C4.5 algorithms to the classification of the severity grades of patients in the ED [48], while Emmanuel et al. used Fuzzy Logic to classify the severity and provide priority for patients [49]. Similarly, Tsai et al. used Long-Short Term Memory neural networks (LSTM) to identify the pain level of patients in emergency triage [50]. In addition, Kijpaisalratana et al. used patients' electronic health records to identify sepsis patients in the ED [51], and Choi et al. utilized the patient data obtained from the ED and XGB model to detect low-risk bacteremia patients [52].

Some researchers have applied machine learning to predict the patient's chief complaints, medical needs, and waiting time in the ED. For instance, Jernite et al. used Support Vector Machine to predict patients' chief complaints at triage time according to the patient's state and nurses' description when they arrive at the emergency department [53]. Sterling et al. used machine learning to predict the required resources at the ED based on the nursing triage notes and the clinical data from the electronic health record (EHR) [54]. Djordje et al. introduced the deep attention model to predict what kinds of medical resources a patient would need when he or she arrives at the emergency department [55]. Furthermore, Ali et al. utilized the Decision Tree (DL) to predict how long patients will stay at the emergency department [56]. Besides, Sterling et al. utilized the NLP techniques of nursing triage notes to predict final emergency department disposition [57].

Despite the potential of data-driven methods demonstrated in emergency care, there is no study of machine learning-supported triage from primary to secondary care. Triaging from primary to secondary care normally involves challenges of heterogeneous data such as structured blood test results and unstructured data like GP referral letters, and real-world GP referral letters are at various levels of details of disease history, drug history, and previous symptoms. There are researches developing machine learning-based clinical outcome forecasting [58] and phenotyping [59] models for patients already diagnosed with rheumatoid arthritis. However, these models target decision support after the patients are diagnosed (after Step 5 in Fig. 1) using Electronic Health Record data generated at the secondary care hospital, for treatment monitoring purposes. There is no machine learning-based method targeting triage referral assessment in Step 3 and 4 of Fig. 1 before a clinic appointment being booked and a final diagnosis being made, when only GP referral letters and blood test results from primary care are available for referral assessment and model training. Moreover, there is no research using the state-of-the-art language model for NLP such as the BERT-based model analyzing GP referral letters. There is no machine learning-based triage referral assessment providing local prediction explanation which is critical in triage decision support practice.

Table 1 summarizes current hospital triage methods and gaps. Our study is the first research in primary care to secondary care triage and addresses the above-mentioned research gaps.

3. Methodology

An overall framework of the proposed heterogeneous data-driven hybrid machine learning model to support triage referral assessment from primary care to secondary care can be found in Fig. 2. Our approach has three sub-models including the Blood Test Result (BTR) model, the General Practitioner Referral Letter (GPRL) model, and the hybrid model.

The GPRL model set out to develop the BERT-based NLP models and select the best-performing one. This model has a series of steps: (a) data cleaning and formatting of original GP referral letters that have been anonymized; (b) text data augmentation methods are used to alleviate the potential data size effect when there are limited machine-readable referral letters; (c) two BERT-based classification models are developed, and threshold optimization is utilized to search for the best classification threshold; (d) the best performing model and parameters are applied to predict the probabilities of diagnosis for new patients.

For the BTR model, the main goal is to determine the proper combination of various missing data imputation methods, sub-sampling methods, and machine learning models. This model also comprises several succeeding steps after the data anonymization: (a) prepare for the BTR training dataset by using different data cleaning methods, including strategies to impute the missing values and to handle the outlier values; (b) different sub-sampling techniques are used; (c) train and validate various machine learning models, and select the best model by tuning the threshold; (d) use the best-combined method to predict the probabilities for new patients.

The third part of our approach is a hybrid model that fuses predictions from GPRL and BTR models if a patient has GPRL data and BTR data available simultaneously. Otherwise, the model will output the classification result directly using the preceding model's result, i.e., either the GPRL model or BTR model depending on data available at the

Table 1
Survey of current hospital triage methods and our study.

Study	Category	Methods	Data Type	Advantages and disadvantages	Triage Application
[15–22,60,61]	EBT	Triage scales, EWS, MEWS, NEWS	Enquiry or scoring	Labor-intensive, manual processes, sometimes may have over triage due to accuracy.	General triage purpose in the ED or sometimes for inpatients.
[26–30]	EBT	Triage scales	Disease-specific enquiry or scoring	Labor intensive, manual processes, used for specific diseases only.	Specific disease triage like obstetrics.
[48,49,53,56]	DDT	NB/C4.5, Fuzzy Logic, SVM, DT	Machine learning model is mainly based on structured data (general patient characteristics, test results, vital signs, etc.)	Mainly used in ED and suitable for predictions during triage where only structured data is available.	Severity grades classification, hospitalization mortality predictions.
Our Approach	DDT	LGBM, GNB, DT, LDA, RF, SVM, BERT, EL	Heterogenous data-driven machine learning methods to address real-world data challenges using transformer-based NLP method and hybrid model to incorporate different data modality	Address healthcare real-world data challenges; can be used for multi-modality data and can be used in different triaging scenarios.	The first data-driven method for triage from primary care to secondary care.

Expert Based Triage (EBT); Data Driven Triage (DDT), Naïve Bayes (NB), LightGBM (LGBM); Gaussian Naïve Bayes (GNB); Decision Tree (DT), Linear Discriminant Analysis (LDA), Random Forest (RF), Support Vector Machine (SVM), Bidirectional Encoder Representations from Transformers (BERT), Ensemble Learning (EL), Early Warning Scores (EWS), Modified Early Warning Score (MEWS), and National Early Warning Score (NEWS).

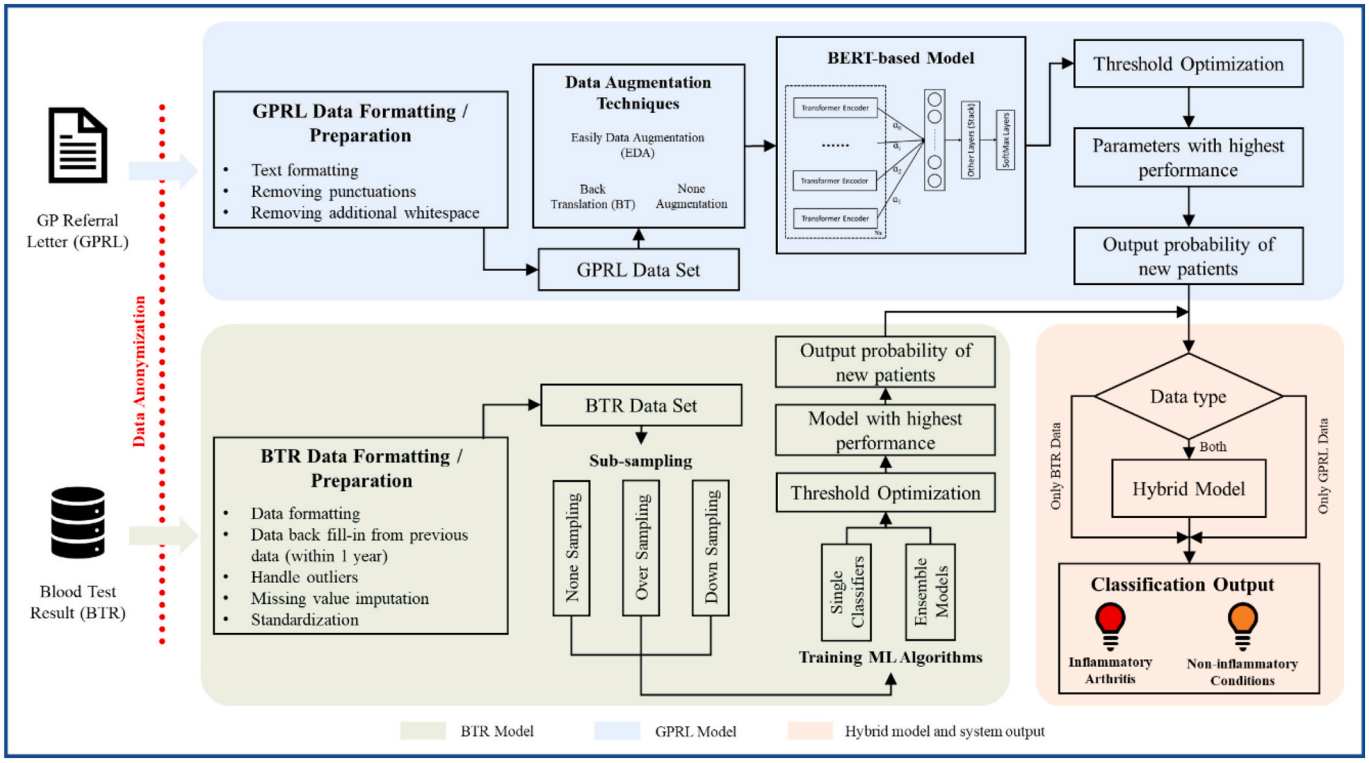


Fig. 2. The framework of the machine learning based triaging methods.

point of triage. The classification result will identify patients with inflammatory arthritis (IA) and patients with non-inflammatory conditions (NIC), which will result in suggesting different triaging routes and clinics in secondary care.

3.1. Methods for GPRL model

3.1.1. Data preparation

GPRL data preparation has two steps: (a) excluding letters without clinical concepts, GPRL letters without any clinical concepts are excluded in our research; (b) removing special characters, which means to format GPRL texts for every patient by removing typical characters like line break character, and extra whitespace character.

3.1.2. Data augmentation

After the data preparation procedure, there are only 332 IA patients in the GPRL dataset, which is relatively small to fine-tune a BERT-based model. To maximize the potential of the BERT-based model in our study, different text augmentation methods are compared with the original dataset, including the Easy Data Augmentation (EDA) [62], and Back Translation (BT) [63], which serve as effective methods for domain applications when the dataset is relatively small.

3.1.3. NLP models for text classification

In this section, we develop different BERT-based models for text classification, which are comprised of the BERT-based model and various classification layers.

3.1.3.1. Multi-layers feature fusion of BERT model (MBERT). Bidirectional Encoder Representations from Transformers (BERT) is the state-of-the-art pre-trained language representation model [64]. According to [65], diverse layers of the BERT are capable of extracting different kinds of features or patterns from the text. For example, the low-layer networks identify phrase-level features, while the intermediate-layer networks extract linguistic features, and the top-layer networks learn semantic features. Thus, with the consideration of the various information extracted by different layers, we weighted-combine different-level hidden layers' outputs as the BERT's output for the following layers, and the weighted parameters for each layer will be learned by the model on the fly. The multi-layers feature fusion method is called the MBERT model. The output of the MBERT model is $O_{MbERT} = \sum_{i=1}^{N}(\alpha_i \bullet h\_o_i)$, where N is the total layers of the BERT model used in our research, i.e., 12 for BERT-base or 24 for BERT-large, and h_o_i is the output of the *i-th* hidden layer of the BERT. The $\alpha_i = Softmax(Dense\_layer(h\_o_i))$ is the weight of the *i-th* layer, and all α_i are summed up to 1, as $\sum_{i=1}^{N}\alpha_i = 1$.

3.1.3.2. Classification models. In our research, we extracted two kinds of text representations for the patient's letter, which are used as the input of the following classification layers, separately including the output of BERT and MBERT models. Based on the different text representations, we add various layers to develop two classification models, including BERT-LL-SL, and MBERT-LL-SL, as illustrated in Fig. 3.

Looking at Fig. 3 (a), BERT-LL-SL, as the basic text classification model in our research, is made up of a BERT model, a fully connected linear layer (LL), and a SoftMax layer (SL). Notably, in this model, we only use the first token vector of the final layer of the BERT model. In Fig. 3 (b), MBERT-LL-SL has the same following layers as BERT-LL-SL except for the BERT model layers. MBERT-LL-SL uses the weighted combination of all hidden layers as the input of the classification layer to predict the labels.

3.2. Methods for BTR model

3.2.1. Data formatting and cleaning

The formatting and cleaning procedure of the BTR dataset consists of two steps: the first step involves the back fill-in of missing data from the recent data. In our research, we use the patient's previous latest records to fill in a part of the missing values. This method is reasonable and acceptable because IA is a long-term disease, and some important blood test indicators might remain stable for a long time according to professional experience and suggestions from clinicians. The second step is to handle outliers. We remove some invalid string values in the data frame, such as "In progress", "Insufficient sample for testing", and so on. Besides, we used one-hot encoding to encode the categorical data like 'negative' and 'positive' values in the BTR data.

3.2.2. Data imputation and sub-sampling

After the data formatting and cleaning process, there is also a tiny percentage of missing values in the BTR dataset, which should be further imputed before feeding to the final machine learning models. In our research, two imputation methods are used, including K-Nearest Neighbors (KNN) imputation and multivariate imputation [66,67]. Because different variables have various data ranges, it is essential to scale the variables to the same data range. In our research, we use standardization to format the dataset. Furthermore, different sub-sampling methods are compared in our research hoping to further boost models' performance, including the no-sampling, random down-sampling, and random up-sampling.

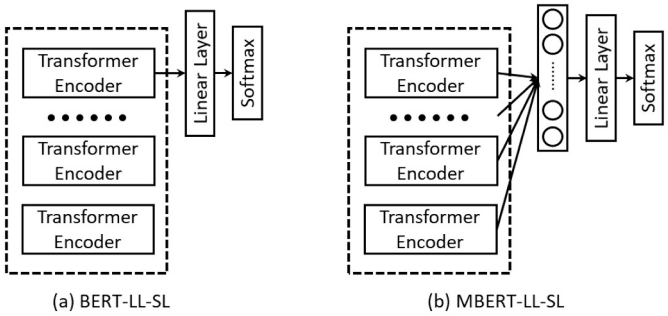


Fig. 3. Text classification models.

3.2.3. Model training and parameters tuning

3.2.3.1. Machine learning models. Generally, different machine learning methods would perform differently on the various datasets based on various data preparation methods, such as missing data imputation and sub-sampling methods. Based on this, in our research, we compare two kinds of machine learning algorithms, including single classifiers and ensemble classifiers. (a) single classifiers, which represent a series of basic machine learning models, such as Gaussian Naïve Bayes (GNB), Decision Tree (DT), Linear Discriminant Analysis (LDA), and Support Vector Machine (SVM); (b) ensemble classifiers, where different base classifiers would be trained and the prediction of the ensemble classifiers is a combination of the outputs of various base classifiers, such as LightGBM (LGBM) and Random Forest (RF).

3.2.3.2. Grid search for optimal parameters. For most machine learning models, there are a plethora of hyper-parameters that would affect the model performance. To unleash the full power of the model, tuning the hyper-parameters is a quintessential procedure to develop a robust as well as practical application. In our research, hyper-parameter space for different models is tuned by using an exhaustive 4-fold grid search and cross-validation scheme.

3.3. Decision fusion and system output

We have thus far developed two models separately for the GPRL dataset and the BTR dataset. As described in Fig. 2, different decision strategies would be applied to predict the final output on the basis of various data types of the patients. First, for patients with either GPRL data or BTR data, the model would directly use the probabilistic prediction of the previous models, i.e., the GPRL model or the BTR model. Second, for patients having both GPRL data and BTR data, we fuse the probabilistic predictions of the GPRL model and the BTR model as the final prediction. Various methods have been proposed to integrate different models' predictions, such as voting, blending, and so on [68,69]. Based on these researches, we proposed a method to fuse probabilistic predictions of heterogeneous data models, as described in Algorithm 1, and tested three kinds of probabilistic fusion methods, including simple average (SAVG), weighted G-Mean (WGM), and weighted AUC (WAUC).

Algorithm 1. Decision fusion and system output method

Input: $\boldsymbol{O}_{\mu}(\boldsymbol{x}) \in Y, Y = \{O_{\mu}: \mu = GPRL, BTR\}$, where $\boldsymbol{O}_{\mu}(\boldsymbol{x})$ is the prediction of μ-*th* model of the patients.

Output: $\boldsymbol{H_F(x)}$, the final output of the decision fusion model.

IF ($\boldsymbol{O_{GPRL}(x)}$ is **not None**) **AND** ($\boldsymbol{O_{BTR}(x)}$ is **None**)

Calculate:

$$\boldsymbol{H_F(x)} = \begin{cases} +1, & O^{+1}_{GPRL}(x) > T_{GPRL} \\ -1, & Otherwise \end{cases}$$

ELSE IF ($\boldsymbol{O_{GPRL}(x)}$ is **None**) **AND** ($\boldsymbol{O_{BTR}(x)}$ is **not None**)

Calculate:

$$\boldsymbol{H_F(x)} = \begin{cases} +1, & O^{+1}_{BTR}(x) > T_{BTR} \\ -1, & Otherwise \end{cases}$$

ELSE

Calculate:

$$\boldsymbol{H_F(x)} = \begin{cases} +1, & \sum_{\mu} \beta_{\mu} O^{+1}_{\mu}(x) > T_E \\ -1, & Otherwise \end{cases}, \forall \mu \in GPRL, BTR$$

ENDIF

RETURN patient's prediction $H_F(x)$.

where T_{GPRL}, T_{BTR}, and T_E are the optimal thresholds for the GPRL model, BTR model, and ensemble model respectively. $O^{+1}_{\mu}(x)$ is predicted probability for positive class. The β_{μ} is the weight of the μ-*th* model. Specifically, $\beta_{\mu} = \frac{1}{2}, \forall \mu \in GPRL, BTR$ for SAVG, $\beta_{\mu} = \frac{G\text{-}Mean_{\mu}}{\sum G\text{-}Mean_{\mu}}, \forall \mu \in GPRL, BTR$ for WGM, and $\beta_{\mu} = \frac{AUC_{\mu}}{\sum_{\mu} AUC_{\mu}}, \forall \mu \in GPRL, BTR$ for WAUC.

3.4. Method for prediction explanation

Deep learning methods including BERT-based models are black-box predictions that cannot be readily explained to clinicians [70]. To benefit from the higher predictive power, it is important to have explainable and transparent DL algorithms for disease classification, as it will help clinicians back-trace disease predictions for transparent triaging recommendations [71]. Local interpretable model-agnostic explanation (LIME) is introduced to investigate the BERT-based model's interpretability in the classification of GPRL texts. Briefly, LIME explains models by using the interpretable algorithm to approximate the predictions of any black-box models [72]. In this paper, we use LIME to identify words and variables that highly contribute to the model's prediction.

3.5. Performance metrics and cross validation

In order to measure the performance of the different previously-mentioned models, we utilize six categories of matrices, including the Accuracy, Precision, Recall, F1-score, G-mean, and AUC. Furthermore, to validate the consistency of the models, the stratified 5-fold cross validation test is applied in this study, which could preserve the percentage of samples for each class. It involves the following steps. First, the dataset is randomly shuffled and evenly split into 5 groups. Second, a unique group is taken out as the test set and the remaining groups as the training set. Third, train the model on the training set and evaluate it on the test set. Fourth, iterate step 2 and step 3, and average the 5-fold cross validation results as the models' final evaluation scores.

4. Experimental results

4.1. Dataset preparation

Our dataset was collected from the Rheumatology Department of a large secondary care hospital in the UK. All data used has been anonymized in accordance with the regulations of data protection and information governance. The whole dataset was split into three sub-datasets according to different data modalities that patients have, including the GPRL dataset, BTR dataset, and GPRL+BTR dataset. The GPRL sub-dataset had 1264 patients referred from February 2018 to July 2021, including 932 NIC patients and 332 IA patients. This dataset was a general natural language description of the patient's physical conditions when they came to the primary care GP clinics, including physical check-ups, simple blood test results, disease history, drug history, GP's consideration, suggestions, etc. The BTR sub-dataset involved 1181 patients with BTR data referred from February 2017 to July 2021, including 353 NIC patients and 828 IA patients. All the input variables of the BTR dataset were categorized into four groups, including patients' demographic information, haematology (routine), blood biochemistry (routine), and immunology. There were only 119 patients who had GPRL data and BTR data simultaneously, which included 65 NIC patients and 54 IA patients, thus we use this sub-dataset (GPRL+BTR Dataset) to test the ensemble method described in Section 3.3.

Different data partition tactics were applied to various datasets. Specifically, for the GPRL model and BTR model, we used 80% to train the model and 20% for the test due to the stratified 5-fold cross validation applied in our study. To validate the ensemble model, we chose the validation: test ratios of 2:1 to split the ensemble dataset, and the validation set is used to calculate the weighted ratios for fusing predictions of the GPRL and BTR models. Furthermore, stratified sampling is used throughout our research to maintain the real distribution of the sub-dataset same as the original dataset.

4.2. Results of GPRL model

4.2.1. Data cleaning and augmentation

Following the processing steps described in Section 3.1.1, we got a dataset of 1264 patients having GPRL data. In order to get the best performance of models, BERT and MBERT described in Section 3.1.3 were separately compared based on different data augmentation

Table 2
Comparison results with 95%-CI of 5-fold cross validation of different GPRL models.

Model	AUG	Precision	Recall	F1-Score	Accuracy	AUC	G-Mean
BERT	Baseline	0.73 ± 0.05	0.69 ± 0.06	0.70 ± 0.06	0.69 ± 0.06	0.69 ± 0.06	0.66 ± 0.06
	EDA	0.69 ± 0.06	0.59 ± 0.06	0.61 ± 0.06	0.59 ± 0.06	0.60 ± 0.06	0.60 ± 0.06
	BT	0.69 ± 0.06	0.61 ± 0.06	0.63 ± 0.06	0.61 ± 0.06	0.62 ± 0.06	0.60 ± 0.06
MBERT	Baseline	**0.79 ± 0.05**	**0.77 ± 0.05**	**0.77 ± 0.05**	**0.77 ± 0.05**	**0.81 ± 0.05**	**0.74 ± 0.05**
	EDA	0.77 ± 0.05	0.74 ± 0.05	0.75 ± 0.05	0.74 ± 0.05	0.80 ± 0.05	0.71 ± 0.06
	BT	0.77 ± 0.05	0.76 ± 0.05	0.76 ± 0.05	0.76 ± 0.05	0.77 ± 0.05	0.70 ± 0.06

Note: Augmentation method (AUG), Original Dataset without augmentation (Baseline), Easily Data Augmentation (EDA), Back Translation Dataset (BT).

Table 3
Summary of the BTR dataset.

Group types	Data sub-items
Demographic information	Age, Gender.
Haematology (Routine)	Haemoglobin, White blood cell count, Platelet count, Red blood cell count, Mean cell volume, Haematocrit, Mean cell haemoglobin, Mean cell haemoglobin conc, Neutrophil count, Lymphocyte count, Monocyte count, Eosinophil count, Basophil count, Erythrocyte sedimentation rate.
Blood biochemistry (Routine)	Sodium, Potassium, Urea level, Creatinine, Albumin, Bilirubin, Alkaline phosphatase, Alanine transaminase, C-reactive protein.
Immunology	Rheumatoid factor, Cyclic citrullinated peptide Ab.

Table 4
Comparison results with 95%-CI of 5-fold cross validation of different BTR machine learning models.

Model	Imp, Sub	Precision	Recall	F1-Score	Accuracy	AUC	G-Mean
GNB	Mul, Under	0.71 ± 0.06	0.64 ± 0.06	0.66 ± 0.06	0.64 ± 0.06	0.70 ± 0.06	0.65 ± 0.06
DT	Mul, Over	0.73 ± 0.06	0.71 ± 0.06	0.72 ± 0.06	0.71 ± 0.06	0.69 ± 0.06	0.68 ± 0.06
LGBM	Mul, No	**0.78 ± 0.05**	**0.77 ± 0.05**	**0.77 ± 0.05**	**0.77 ± 0.05**	**0.81 ± 0.05**	**0.73 ± 0.06**
LDA	KNN, No	0.73 ± 0.06	0.66 ± 0.06	0.68 ± 0.06	0.66 ± 0.06	0.71 ± 0.06	0.67 ± 0.06
RF	Mul, No	0.76 ± 0.05	0.73 ± 0.06	0.74 ± 0.06	0.73 ± 0.06	0.79 ± 0.05	0.72 ± 0.06
SVM	Mul, Over	0.72 ± 0.06	0.68 ± 0.06	0.69 ± 0.06	0.68 ± 0.06	0.73 ± 0.06	0.67 ± 0.06

Note: Gaussian Naïve Bayes (GNB), Decision Tree (DT), LightGBM (LGBM), Linear Discriminant Analysis (LDA), Random Forest (RF), Support Vector Machine (SVM), Multivariate (Mul), Sub-sampling method (Sub), Imputation method (Imp).

methods, including the original dataset (Baseline), EDA-augmented dataset (EDA), and BT-augmented dataset (BT), as described in Section 3.1.2. For the EDA method, we generated approximately four augmented sentences per original sentence. For BT augmentation, we used a pre-trained translation package provided by Transformers to translate the original training dataset to four kinds of languages (French, Spanish, Romanian, and Romansh), and then translate them back to English to get a new training dataset.

4.2.2. Comparison study of GPRL models

A comparison study of the BERT and MBERT models is described in Table 2, which is based on various data augmentation methods as described in Section 4.2.1. Overall, the MBERT model achieved the best performance in all measures compared with the BERT model, and specifically achieving the weighted precision, recall, and F1-Score of 0.79, 0.77, and 0.77.

The MBERT model trained on the dataset without augmentation (Baseline) achieves the best performance. Specifically, it achieves the Accuracy, AUC, and G-Mean of 0.77, 0.81, and 0.74, as well as weighted values of Precision, Recall, and F1-Score of 0.79, 0.77, and 0.77. It is clear that data augmentation methods such as EDA and BT cannot improve the model's performance. Thus, the baseline MBERT model trained on the original data set will be used for triage classification.

Table 5
Comparison of ensemble model results with 95%-CI.

Method	Precision	Recall	F1-Score	Accuracy	AUC	G-Mean
SBTR	0.73 ± 0.14	0.72 ± 0.14	0.73 ± 0.14	0.73 ± 0.14	0.81 ± 0.12	0.72 ± 0.14
SGPRL	0.81 ± 0.12	0.80 ± 0.12	0.80 ± 0.12	0.80 ± 0.12	0.79 ± 0.13	0.78 ± 0.13
WAUC	**0.83 ± 0.12**	**0.82 ± 0.12**	**0.83 ± 0.12**	**0.82 ± 0.12**	**0.90 ± 0.09**	**0.83 ± 0.12**

Note: Single BTR(SBTR), Single GPRL (SGPRL), Weighted AUC (WAUC).

4.3. Results of BTR model

4.3.1. Data cleaning and preparation

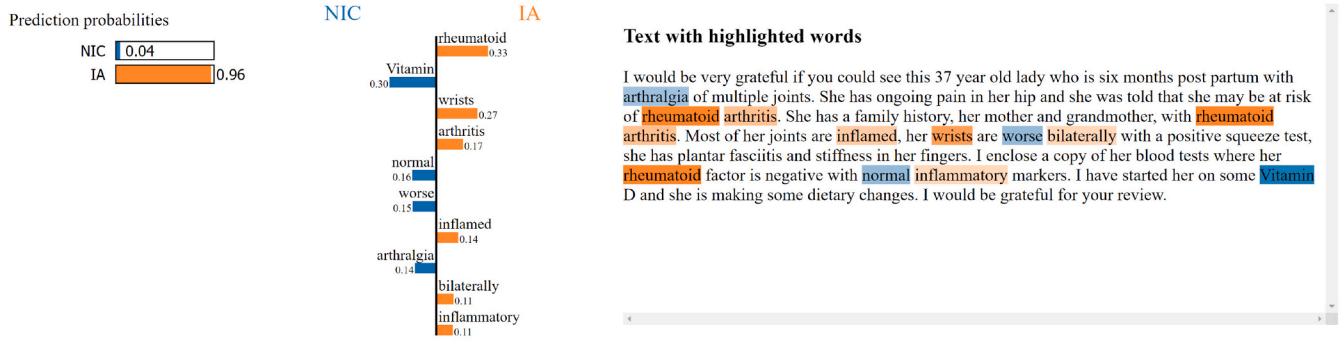
According to the feature selection method described in Section 3.2.1, a total number of 27 blood test result features in the category of demographic information, haematology (routine), blood biochemistry (routine) and immunology were used to train the model, as shown in Table 3. After feature selection, 1181 patients have BTR data in total.

4.3.2. Comparison study of BTR models

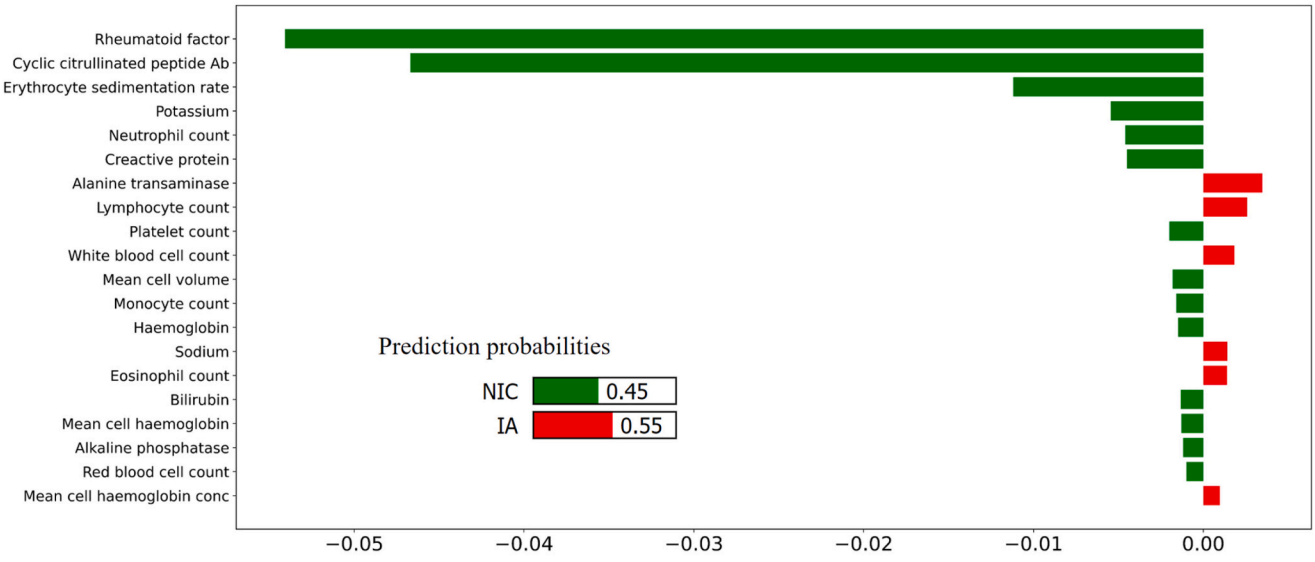
We conducted 5-fold cross validation to test the consistency of the results and exhaustive experiments to search for the best combination of the different missing data imputation methods (KNN and Multivariate) and the various sub-sampling methods (No sampling, Over-sampling, and Under-sampling). Table 4 indicates the best 5-fold cross validation results of six various machine learning models depending on the missing data imputation method and the sub-sampling method. It was apparent from Table 4 that LGBM consistently outperforms the other five models in the 5-fold cross-validation test, and the Accuracy, AUC, and G-Mean values were 0.77, 0.81, 0.73, using the multivariate imputation and no sub-sampling method. Notably, our model resulted in the weighted precision, recall, F1-Score values of 0.78, 0.77, and 0.77 for identifying IA and NIC. Therefore, the LGBM model with the multivariate imputation and the no sub-sampling method will be used for triage classification.

4.4. Results of hybrid model

The GPRL+BTR dataset was split into the validation and test set with a ratio of 2:1. We used the validation set to calculate the weights for different models and verified them on the testing dataset. The weights of



(a) LIME output of GPRL data.



(b) LIME output of BTR data.

Fig. 4. Example of LIME prediction explanation of a triage recommendation of inflammatory arthritis (ensembled risk probabilities: NIC: 0.27 and IA: 0.73)

GPRL and BTR models for best combination are calculated dynamically in each model training and updating rather than once. Same as parameters in GPRL and BTR, they will be updated if we retrain the model with new referral data.

Table 5 shows the results of ensemble models on the GPRL+BTR dataset. Overall, WAUC achieved the best Accuracy, AUC, and G-Mean values of 0.82, 0.90, and 0.83, which were significantly better than single GPRL and BTR models. Furthermore, the SAVG and WGM have the same results as WAUC. Specifically, our hybrid model achieves the weighted values of Precision, Recall, and F1-Score of 0.83, 0.82, and 0.83.

4.5. *Explaining disease predictions for patient triaging*

As discussed in Section 3.4, we have developed local prediction explanation into our triaging classification processes using the LIME method. LIME provides explanations in the form of highlighting the words and blood test results that are more important for the model prediction of a patient having inflammatory arthritis or other inflammatory conditions, from a patient's input GP referral letter text and blood testing data. The explainable triaging classification could help clinicians to make a final decision about triaging a patient to IA or NIC clinics, help build trust in the triaging model, and also serve as a confirmation that the model-internal logic is sound and reliable.

As shown in Fig. 4, we exemplify the LIME explanations for triaging recommendation of a real IA patient estimated to have IA with 0.73 probability by the hybrid WAUC model. From Fig. 4(a), it was evident that the "arthralgia", "rheumatoid arthritis", "inflamed", "wrists", and so on in the text were of relatively-high importance to identifying a patient to have inflammatory arthritis from the GP referral letter. From this patient's blood test results, test results like "Alanine transaminase", "Lymphocyte count", and so on had a high positive contribution to the classification as having inflammatory arthritis, as shown in Fig. 4(b).

5. Real-world case study and evaluation

5.1. *Practical implications on triage referral assessment pathway*

In real-life practice, our model offers a readily-available means of rapid streamlined referral assessment for clinicians to identify patients with suspected IA or NIC, on the basis of their GPRL and BTR data. Patients identified with suspected IA will be booked to an outpatient IA clinic to confirm the final diagnosis and pathway. If the patient is diagnosed as IA in the outpatient clinic, they will be offered early

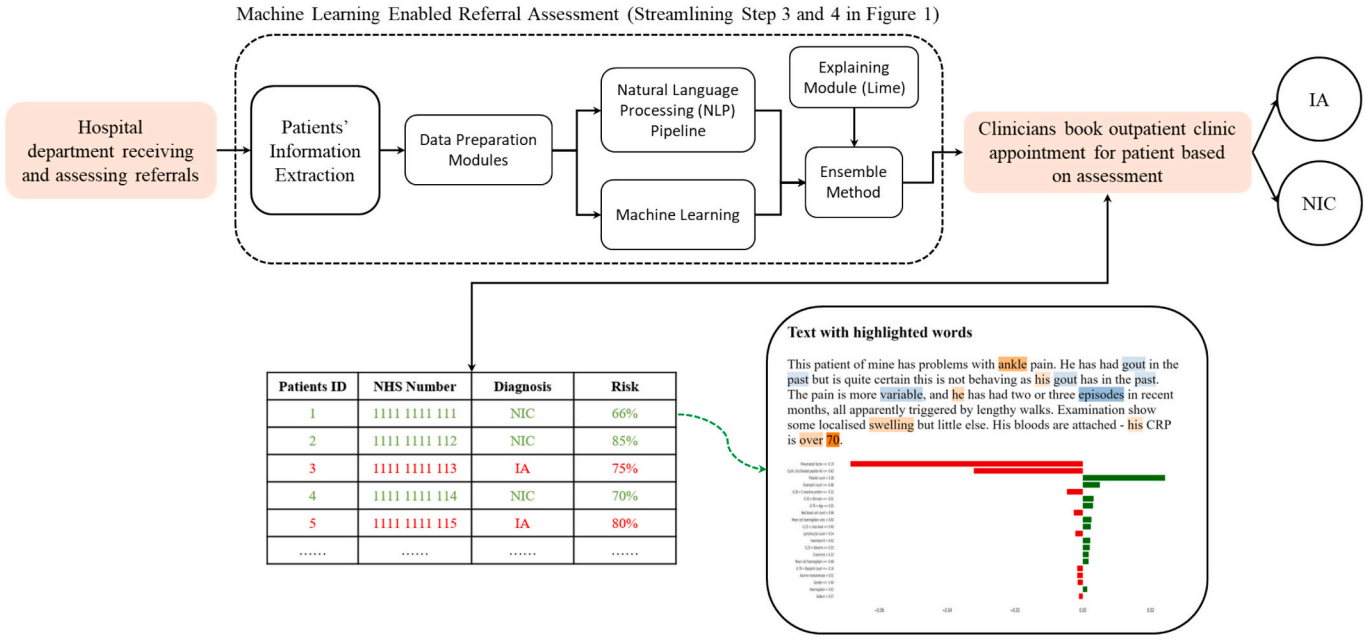


Fig. 5. Hybrid machine learning supported referral assessment process. Inflammatory Arthritis (IA), Non-inflammatory Condition (NIC).

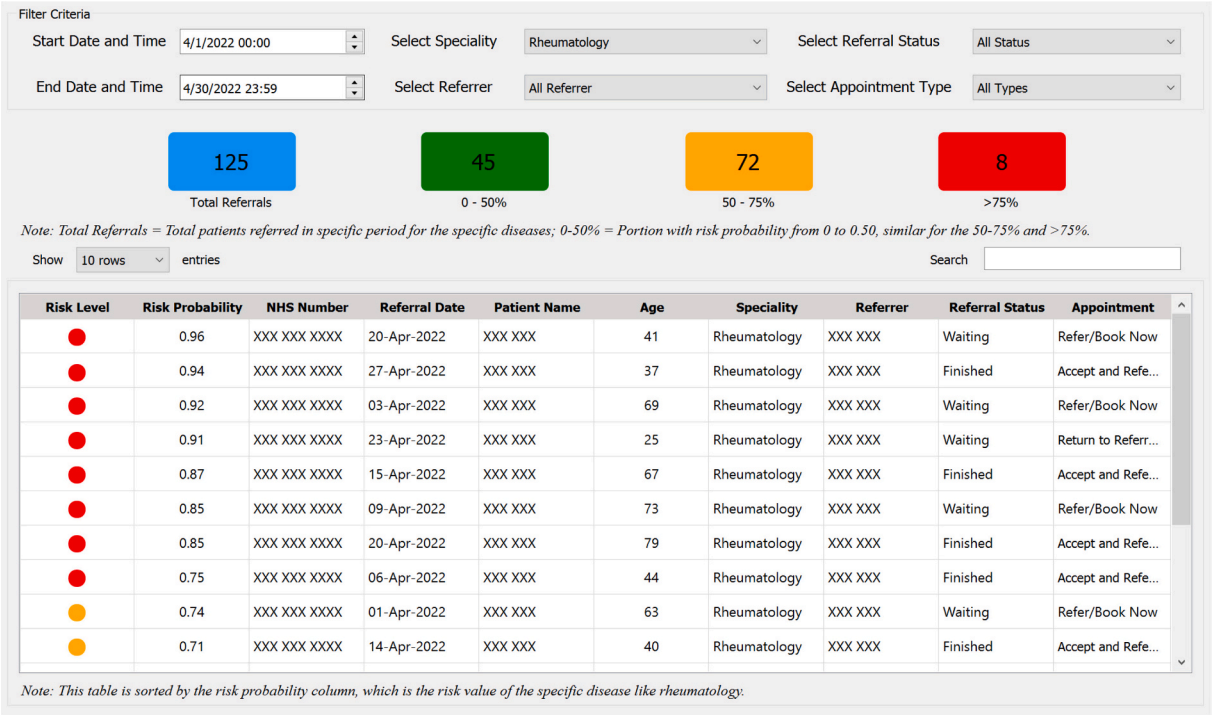


Fig. 6. Demonstration: real-time risk stratification of referrals waiting to be triaged through the decision support system.

treatment with disease-modifying anti-inflammatory drugs such as methotrexate or biologic drugs directed at specific inflammatory cytokines [3,73–75].

From Fig. 1, a typical current referral pathway from a GP to a rheumatologist at the secondary care hospital in the UK is for the GP to see the patient and write a referral letter to a rheumatologist. The patient then selects a rheumatology department on the NHS e-Referral Service (formerly known as Choose and Book) website. The GP referral letter is directed by the e-Referral Service to the selected secondary care hospital department for a rheumatologist to assess the referral and triage the patient to a suitable clinic. This takes 2 h a day screening and answering GP referrals (10 h a week) which equals 2.5 clinics, and has to be fitted into time devoted to other clinical activities. Streamlined assessment of referrals would reduce the delay due to human triage and speed up the e-Referral process at all stages. This will also release the time of rheumatologists to see patients as clinical time currently used for triage of the many patients referred each week, will be regained.

As shown in Fig. 5, our referral assessment tool will be used to analyze heterogenous data automatically. The clinicians (consultants) will be presented with the risk of having IA or NIC and prioritize the

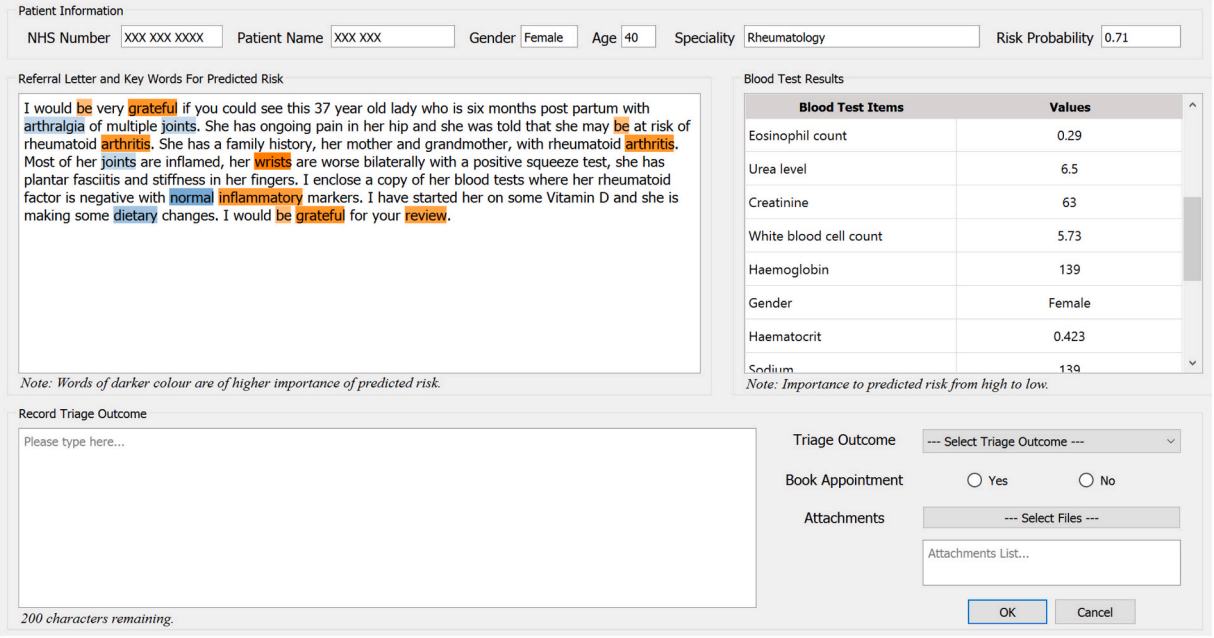


Fig. 7. Demonstration: key information that contributes to the risk stratification of individual referrals.

clinics according to the risk. The prediction risk is explainable for each patient from both GP referral letters and blood test results with key symptoms and test results highlighted by the LIME method. We co-developed the decision support tool prototype based on our model with clinicians in a large secondary care hospital in the UK and the pilot details are discussed in Section 5.2.

5.2. Triage referral assessment decision support system pilot

The triage referral assessment decision support tool is being co-developed with the clinician team and is under the preparation of rolling out in the rheumatology department of a secondary health care hospital in the UK. We developed an easy-to-use software demonstration dedicated to simplifying the referral assessment processes for clinicians at the secondary care hospital. Furthermore, all models in this system will be auto-updated periodically when more new referral data are available, including the GPRL, BTR, and hybrid models.

As shown in Fig. 5 and Fig. 6, instead of clinicians searching for the individual patient's referral letter and blood test results manually and separately, our tool will extract data automatically from different sources for real-time risk stratification.

The referral triage assessment decision support system features real-time risk stratification of patients having inflammatory arthritis or non-inflammatory conditions, at the individual patient level and the point of referral assessment. Patient referrals are sorted from high to low-risk probabilities of having IA, as well as high (red), medium (yellow), and low risk (green) groups. Fig. 6 shows the screenshot of the decision support tool in the real-world application.

During referral assessment, clinicians will first look at the risk stratifications of all referrals waiting to be triaged in Fig. 6, and then they can click to investigate individual referrals for risk stratification details as shown in Fig. 7. Take a referral with a predicted 71% probability of having IA as an example, the system visualizes important words in the GP referral letter that contribute to the risk prediction (left part of Fig. 7), and rank blood test result according to their importance to the predicted risk from high to low (right part of Fig. 7). Risk stratification and explanations of the risk in Fig. 6 and Fig. 7 will provide decision support for clinicians to make decision on triage i.e., book outpatient IA appointment or triage to other pathways.

Table 6
Comparison results with 95%-CI of human and model performance.

Types	Precision	Recall	F1-Score	Accuracy	G-Mean
Clinician	0.80 ± 0.08	0.78 ± 0.09	0.79 ± 0.09	0.78 ± 0.09	0.77 ± 0.09
Model	0.83 ± 0.08	0.81 ± 0.08	0.81 ± 0.08	0.81 ± 0.08	0.81 ± 0.08

5.3. Decision support pilot evaluation

To evaluate the practical value of our model, a real-world pilot was conducted in June and July 2022 at the same hospital described in Section 4.1. In this real-world pilot, we designed a human versus machine trial to compare the practical performance of our model and clinicians in terms of referral assessment accuracy and time spent in referral assessment. We also collected qualitative feedback from clinicians who used the pilot decision support tool in referral assessment.

5.3.1. Referral assessment accuracy

To compare the referral assessment accuracy of the model and clinicians, we collected referral and diagnosis data of 88 patients referred to the hospital from November 2021 to December 2021. We chose this period because it takes around 3 months for the final coded diagnosis information to be available in the Electronic Health Record after the outpatient clinic appointment because time is needed for the coding team to finalize the ICD10 code according to clinical notes in the outpatient appointment. It takes a longer time in 2021/22 due to the appointments backlog caused by the COVID-19 pandemic. The coded diagnosis information is used as the benchmark to calculate the precision, recall, and accuracy of clinicians' assessment and the triage decision, as well as the model's prediction. According to the diagnosis information, there are 62 patients diagnosed as NIC and 26 patients diagnosed as IA. Furthermore, to calculate clinicians' accuracy of referral assessment and triage decision, four metrics (false positive, false negative, true positive, and true negative) are defined to calculate the confusion matrix based on the assessment/triage outcome and the final diagnosis.

Table 6 shows that our model outperforms clinicians in the values of all measure metrics. Our model achieves the Accuracy and G-Mean of

0.81 and 0.81 compared with the 0.78 and 0.77 by physicians. Specifically, it achieves the weighted Precision, Recall, and F1-Score of 0.83, 0.81, and 0.81, which are clearly better than clinicians. This means more patients will be booked to the right clinic appointments faster with the support of our model, i.e., the patients will not need to attend unnecessary appointments back and forth before they are booked for the right clinics, thus reducing the referral to treatment time.

5.3.2. Time spent in referral assessment

The estimated time for referral assessment was compared between a clinician assessing without decision support in June 2022 and a clinician assessing with decision support of our machine learning model in July 2022. There are approximately 84 referrals per month (around 1000 per year). For the clinician assessing without decision support, 2 h per day (10 h per week) is needed to go through referral data and suggest clinic appointment that patient to be triaged, while the model predicts disease probability and visualize the explainable results in real-time and it is expected that only 2 h per week will be needed to to assess the referrals with the decision support.

The pilot shows our model can potentially reduce clinicians' time spent on referral assessment from 10 h per week (equals to 2.5 clinics which have to be fitted into time devoted to other clinical activities), to 2 h per week. This means 8 h (equals to 2 clinics) can be saved for other clinical activities and for clinicians to see more patients and provide better care. This also means faster referral assessment and less delay between a hospital rheumatology department receiving a referral from GP for suspected EIA and the date of clinic assessment for diagnosis. This also contributes to early diagnosis and treatment that are critical to preventing patient distress, serious complications, avoidable disability, and potential loss of employment and quality of life.

5.3.3. End users' feedback

Positive feedback from clinical end users of the referral assessment decision support system has been received in the trial. According to the feedback, the model is gaining more trust than the traditional approach because 1) it demonstrated better performance than humans and all the existing clinical criteria. For example, existing clinical criteria of ACR/EULAR 2010 has 0.74 recall (sensitivity) and specificity 0.66 and 0.79 AUC [76]; 2) we are not automating the referral assessment using machine learning. Instead, our model will provide decision support information of disease risk probability and highlight important words in the GP referral letters and blood test results that contributed to the predicted risk. This provides transparency and explainability of the underlying logic and ensures that the model can be checked for the reliability of model recommendations. It is reported that the explainable prediction information shows similar ways of doctors assessing the referrals thus further improving the confidence of using the model.; 3) The risk stratification and explainable decision support also provide clinicians an intuitive and straightforward way to prioritize high-risk patients and identify key information faster. The process of referral for specialist rheumatology assessment from the moment of referral to being seen in an appropriate clinic would be shortened from two to three weeks to a few days. This matters to the patients themselves and would help in achieving the key target required by NICE Quality Standard 33 of getting patients with IA onto disease-modifying drugs within 6 weeks of referral [77].

6. Conclusion and discussion

6.1. Summary of results

Inflammatory arthritis (IA) is an autoimmune disease that can cause severe joint damage and disabilities. Accurate and fast referral assessment of IA for triaging is an important but challenging task due to vague symptoms and manual processes. In this research, a heterogeneous data-driven hybrid machine learning approach was developed to perform the IA and non-inflammatory conditions (NIC) disease triage from primary care to secondary care. Specifically, a Multi-layers Feature Fusion BERT model (MBERT) was developed, and a comparison study between BERT and MBERT has been carried out to classify IA and NIC triages from GP referral letters. Moreover, various data augmentations including EDA and BT have been compared with NLP models in our study. We also developed an ensemble method to fuse the probabilistic predictions from multimodal data including natural language in the GP referral letters and blood test results, which can predict whether it is needed to be triaged for IA and NIC with weighted values of precision, recall, and F1-Score of 0.83, 0.82, and 0.83, and accuracy, AUC and G-Mean of 0.82, 0.90, and 0.83. Furthermore, a real-world case study showed that our model achieved the weighted precision, recall, F1-Score of 0.83, 0.81, 0.81, and accuracy 0.81, comparing the weighted precision, recall, F1-Score of 0.80, 0.78, 0.79, and accuracy 0.78 of clinicians in the trial. The pilot also shows that our model enabled decision support can save clinicians 8 h per week in assessing the referral assessment. Our approach also has superior diagnostic performance and reliability than existing clinical criteria for IA [76], which can be used in cohorts with diverse disease manifestations and can be adapted for primary to secondary care triaging in other diseases.

6.2. Limitations and future research

Our study lays the groundwork for future research upon several aspects. Apart from GP referral letters and blood testing, we will collect electronic Patient Reported Outcomes (ePROMs) and remote blood testing monitoring data for clinical phenotyping to map different phenotypes with different diagnoses so that personalized treatment recommendations could be provided at the time of triage. We will also continue our real-world pilot and human versus machine trial in the live referral triage process to further test the model's reliability and effectiveness. If the model proves to be effective in a longer pilot consistently, we will further extend our triaging model into other specialties that are under pressure with demand often outstripping clinical capacity such as gastroenterology and cardiology, etc. Finally, our triage tool is dependent on GPRL and BTR data being machine-readable, which is currently variable with the digitalization levels of the hospitals. However, with the digital transformation and NHS digital-first strategy, this tool will make a great impact to save clinicians time and improve efficiency.

CRediT authorship contribution statement

Bing Wang: Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Visualization. **Weizi Li:** Conceptualization, Methodology, Investigation, Resources, Writing – review & editing, Supervision, Project administration, Funding acquisition. **Anthony Bradlow:** Conceptualization, Methodology, Validation, Investigation, Resources, Writing – review & editing, Project administration, Funding acquisition. **Eghosa Bazuaye:** Conceptualization, Software, Resources, Data curation, Project administration, Funding acquisition. **Antoni T.Y. Chan:** Conceptualization, Methodology, Validation, Investigation, Resources, Writing – review & editing, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgment

This work is funded by Health Innovation Partnership fund of Royal Berkshire NHS Foundation Trust and University of Reading; UK Engineering and Physical Sciences Research Council (EPSRC, grant number EP/W000652/1); Economic and Social Science Research Council (ESRC, grant number ES/S501785/1). We thank the High Performance Computation (HPC) services provided by the Reading Academic Computing Cluster (RACC).

References

[1] L. Kay, P. Lanyon, A. MacGregor, Rheumatology GIRFT Programme National Specialty Report, 2021.
[2] R.J. Stack, et al., Symptom complexes at the earliest phases of rheumatoid arthritis: a synthesis of the qualitative literature, Arthritis Care Res. 65 (12) (2013) 1916–1926.
[3] NRAS, The National Early Inflammatory Arthritis Audit (NEIAA), vol. Second Annual Report, National Rheumatoid Arthritis Society, 2021.
[4] N. Digital, Referral Assessment Services - NHS e-Referral Service, NHS Digital, 2022.
[5] C. Foot, C. Naylor, C. Imison, The Quality of GP Diagnosis and Referral, 2010.
[6] RCGP, Quality Patient Referrals - Right Service, Right Time, Royal College of General Practitioners, 2018.
[7] W.S. Hong, A.D. Haimovich, R.A. Taylor, Predicting hospital admission at emergency department triage using machine learning, PLoS One 13 (7) (2018), e0201016.
[8] S. Swaminathan, et al., A machine learning approach to triaging patients with chronic obstructive pulmonary disease, PLoS One 12 (11) (2017), e0188532.
[9] R. Miotto, et al., Deep patient: an unsupervised representation to predict the future of patients from the electronic health records, Sci. Rep. 6 (1) (2016) 26094, 2016/05/17.
[10] E.R. Weinerman, et al., Yale studies in ambulatory medical care. V. Determinants of use of hospital emergency services, Am. J. Publ. Health Nationss Health 56 (7) (1966) 1037–1056.
[11] R. Sánchez-Salmerón, et al., Machine learning methods applied to triage in emergency services: a systematic review, Int. Emerg. Nurs. 60 (2022), 101109.
[12] L. Moxham, K. McMahon-Parkes, An evaluation of the impact of advanced nurse practitioner triage and clinical intervention for medically expected patients referred to an acute National Health Service hospital, J. Clin. Nurs. 29 (19–20) (2020) 3679–3686.
[13] A. Eccles, et al., Patient use of an online triage platform: a mixed-methods retrospective exploration in UK primary care, Br. J. Gen. Pract. 69 (682) (2019) e336–e344.
[14] S. Rushton, et al., Effectiveness of Remote Triage: A Systematic Review, 2020.
[15] G. FitzGerald, et al., Emergency department triage revisited, Emerg. Med. J. 27 (2) (2010) 86–92.
[16] A. Hodge, et al., A review of the quality assurance processes for the Australasian triage scale (ATS) and implications for future practice, Australas. Emerg. Nurs. J. 16 (1) (2013) 21–29.
[17] K. Mackway-Jones, J. Marsden, J. Windle, Emergency Triage: Manchester Triage Group, John Wiley & Sons, 2013.
[18] R. Beveridge, CAEP issues. The Canadian triage and acuity scale: a new and critical element in health care reform. Canadian Association of Emergency Physicians, J. Emerg. Med. 16 (3) (May-Jun 1998) 507–511.
[19] D.R. Eitel, et al., The emergency severity index triage algorithm version 2 is reliable and valid, Acad. Emerg. Med. 10 (10) (2003) 1070–1080.
[20] R. Morgan, F. Williams, M. Wright, An early warning scoring system for detecting developing critical illness, Clin. Intens. Care 8 (2) (1997) 100.
[21] C.P. Subbe, et al., Validation of a modified early warning score in medical admissions, Qjm 94 (10) (2001) 521–526.
[22] G.B. Smith, et al., The ability of the National Early Warning Score (NEWS) to discriminate patients at risk of early cardiac arrest, unanticipated intensive care unit admission, and death, Resuscitation 84 (4) (2013) 465–470.
[23] S. O'Neill, et al., Why do healthcare professionals fail to escalate as per the early warning system (EWS) protocol? A qualitative evidence synthesis of the barriers and facilitators of escalation, BMC Emerg. Med. 21 (1) (2021) 1–19.
[24] S. Finnikin, V. Wilke, "What's behind the NEWS? National Early Warning Scores in primary care," 695, Br. J. Gen. Pract. (2020) 272–273.
[25] PULSE, GP clinical judgement leads to 20% fewer referrals than NEWS score, finds study, 28 August 2022. https://www.pulsetoday.co.uk/news/uncategorised/gp-clinical-judgement-leads-to-20-fewer-referrals-than-news-score-finds-study/.
[26] C. Ruhl, et al., Content validity testing of the maternal fetal triage index, J. Obstet. Gynecol. Neonatal. Nurs. 44 (6) (2015) 701–709.
[27] D.S. Smithson, et al., Implementing an obstetric triage acuity scale: interrater reliability and patient flow analysis, Am. J. Obstet. Gynecol. 209 (4) (2013) 287–293.
[28] S. Kenyon, et al., The design and implementation of an obstetric triage system for unscheduled pregnancy related attendances: a mixed methods evaluation, BMC Pregnan. Childbirth 17 (1) (2017) 1–10.
[29] A. Moudi, et al., The development and validation of an obstetric triage acuity index: a mixed-method study, J. Matern. Fetal Neonatal Med. (2020) 1–11.
[30] N. Veit-Rubin, et al., Validation of an emergency triage scale for obstetrics and gynaecology: a prospective study, BJOG Int. J. Obstet. Gynaecol. 124 (12) (2017) 1867–1873.
[31] D. Aletaha, J.S. Smolen, Diagnosis and Management of Rheumatoid Arthritis: a review, JAMA 320 (13) (2018) 1360–1372.
[32] NHS, Rheumatoid arthritis, 28 August 2022. https://www.nhs.uk/conditions/rheumatoid-arthritis/diagnosis/.
[33] M. Clinic, Rheumatoid Arthritis. https://www.mayoclinic.org/diseases-conditions/rheumatoid-arthritis/diagnosis-treatment/drc-20353653, 28 August 2022.
[34] R. C. o. G. Practitioners, The 2022 GP: A Vision for General Practice in the Future NHS, Royal College of General Practitioners, 2013.
[35] V. Mahajan, T. Singh, C. Azad, Using telemedicine during the COVID-19 pandemic, Indian Pediatr. 57 (7) (2020) 658–661.
[36] F.R. Hobbs, et al., Clinical workload in UK primary care: a retrospective analysis of 100 million consultations in England, 2007–14, Lancet 387 (10035) (2016) 2323–2330.
[37] E. Fletcher, et al., Quitting patient care and career break intentions among general practitioners in south West England: findings of a census survey of general practitioners, BMJ Open 7 (4) (2017), e015853.
[38] J.W. Joseph, et al., Deep-learning approaches to identify critically ill patients at emergency department triage using limited information, J. Am. Coll. Emerg. Phys. Open 1 (5) (2020) 773–781.
[39] M. Klug, et al., A gradient boosting machine learning model for predicting early mortality in the emergency department triage: devising a nine-point triage score, J. Gen. Intern. Med. 35 (1) (2020) 220–227.
[40] J.-M. Kwon, et al., Deep learning algorithm to predict need for critical care in pediatric emergency departments, Pediatr. Emerg. Care 37 (12) (2021) e988–e994.
[41] Y. Raita, et al., Emergency department triage prediction of clinical outcomes using machine learning models, Crit. Care 23 (1) (2019) 1–13.
[42] J.Y. Yu, et al., Machine learning and initial nursing assessment-based triage system for emergency department, Healthc. Inform. Res. 26 (1) (2020) 13–19.
[43] S. Levin, et al., Machine-learning-based electronic triage more accurately differentiates patients with respect to clinical outcomes compared with the emergency severity index, Ann. Emerg. Med. 71 (5) (2018) 565–574, e2.
[44] D.J. Teubner, et al., Model to predict inpatient mortality from information gathered at presentation to an emergency department: the triage information mortality model (TIMM), Emerg. Med. Aust. 27 (4) (2015) 300–306.
[45] A. Zlotnik, et al., Building a decision support system for inpatient admission prediction with the Manchester triage system and administrative check-in variables, CIN Comput. Inform. Nurs. 34 (5) (2016) 224–230.
[46] É. Arnaud, et al., Deep learning to predict hospitalization at triage: Integration of structured data and unstructured text, 2022, pp. 4836–4841.
[47] B. Tahayori, N. Chini-Foroush, H. Akhlaghi, Advanced natural language processing technique to predict patient disposition based on emergency triage notes, Emerg. Med. Aust. 33 (3) (2021) 480–484.
[48] D. Zmiri, Y. Shahar, M. Taieb-Maimon, Classification of patients by severity grades during triage in the emergency department using data mining methods, J. Eval. Clin. Pract. 18 (2) (2012) 378–388.
[49] E.S. Sánchez Velarde, et al., Fuzzy-state machine for Triage priority classifier in emergency room, 2022, pp. 1488–1491.
[50] F.-S. Tsai, et al., Embedding stacked bottleneck vocal features in a LSTM architecture for automatic pain level classification during emergency triage, 2022, pp. 313–318.
[51] N. Kijpaisalratana, et al., Machine learning algorithms for early sepsis detection in the emergency department: a retrospective study, Int. J. Med. Inform. 160 (2022), 104689.
[52] D.H. Choi, et al., Prediction of bacteremia at the emergency department during triage and disposition stages using machine learning models, Am. J. Emerg. Med. 53 (2022).
[53] Y. Jernite, et al., Predicting chief complaints at triage time in the emergency department, 2022.
[54] N.W. Sterling, et al., Prediction of emergency department resource requirements during triage: an application of current natural language processing techniques, J. Am. Coll. Emerg. Phys. Open 1 (6) (2020) 1676–1683.
[55] D. Gligorijevic, et al., Deep attention model for triage of emergency department patients, 2022, pp. 297–305.
[56] A. Azari, V.P. Janeja, S. Levin, Imbalanced learning to predict long stay Emergency Department patients, 2022, pp. 807–814.
[57] N.W. Sterling, et al., Prediction of emergency department patient disposition based on natural language processing of triage notes, Int. J. Med. Inform. 129 (2019) 184–188.
[58] B. Norgeot, et al., Assessment of a deep learning model based on electronic health record data to forecast clinical outcomes in patients with rheumatoid arthritis, JAMA Netw. Open 2 (3) (2019) e190606.
[59] R.J. Carroll, et al., Portability of an algorithm to identify rheumatoid arthritis in electronic health records, J. Am. Med. Inform. Assoc. 19 (e1) (2012) e162–e169.
[60] J. Roukema, et al., Validity of the Manchester triage system in paediatric emergency care, Emerg. Med. J. 23 (12) (2006) 906–910.
[61] S. Gerry, et al., Early warning scores for detecting deterioration in adult hospital patients: systematic review and critical appraisal of methodology, BMJ 369 (2020).
[62] J. Wei, K. Zou, EDA: easy data augmentation techniques for boosting performance on text classification tasks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2022, pp. 6382–6388.

[63] R. Sennrich, B. Haddow, A. Birch, Improving neural machine translation models with monolingual data, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2022, pp. 86–96.
[64] J. Devlin, et al., BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2022, pp. 4171–4186.
[65] G. Jawahar, B. Sagot, D. Seddah, What does BERT learn about the structure of language?, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2022, pp. 3651–3657.
[66] O. Troyanskaya, et al., Missing value estimation methods for DNA microarrays, Bioinformatics 17 (6) (2001) 520–525.
[67] S. Van Buuren, K. Groothuis-Oudshoorn, Mice: multivariate imputation by chained equations in R, J. Stat. Softw. 45 (2011) 1–67.
[68] M. Ganaie, M. Hu, Ensemble deep learning: a review, arXiv preprint arXiv: 2104.02395, 2021.
[69] Y. Feng, X. Wang, J. Zhang, A heterogeneous ensemble learning method for neuroblastoma survival prediction, IEEE J. Biomed. Health Inform. 26 (4) (2021) 1472–1483.
[70] S.M. Lauritsen, et al., Explainable artificial intelligence model to predict acute critical illness from electronic health records, Nat. Commun. 11 (1) (2020) 3852, 2020/07/31.
[71] E.J. Topol, High-performance medicine: the convergence of human and artificial intelligence, Nat. Med. 25 (1) (2019) 44–56.
[72] M.T. Ribeiro, S. Singh, C. Guestrin, Why should i trust you?, in: Explaining the Predictions of Any Classifier, 2022, pp. 1135–1144.
[73] M.A. Bukhari, et al., Influence of disease-modifying therapy on radiographic outcome in inflammatory polyarthritis at five years: results from a large observational inception study, Arthritis Rheum. 48 (1) (Jan 2003) 46–53.
[74] V. Nell, et al., Benefit of very early referral and very early therapy with disease-modifying anti-rheumatic drugs in patients with early rheumatoid arthritis, Rheumatology 43 (7) (2004) 906–914.
[75] N.J. Wiles, et al., Reduced disability at five years with early treatment of inflammatory polyarthritis: results from a large observational cohort, using propensity models to adjust for disease severity, Arthritis Rheum. 44 (5) (May, 2001) 1033–1042.
[76] C. Alves, et al., Diagnostic performance of the ACR/EULAR 2010 criteria for rheumatoid arthritis and two diagnostic algorithms in an early arthritis clinic (REACH), Ann. Rheum. Dis. 70 (9) (2011) 1645–1647.
[77] Nice, Rheumatoid arthritis in over 16s, National Institute for Health and Care Excellence, 2020.

Bing Wang is currently a PhD candidate in informatics and system science at the Informatics Research Center, Henley Business School, University of Reading. His research interests are Natural Language Processing, Machine Learning and Graph Machine Learning. He has been working as a data scientist at Royal Berkshire NHS Foundation Trust since December 2019 during his PhD. He obtained his master's degree in engineering at Xi'an Jiaotong University.

Weizi Li is a Professor of Informatics and Digital Health, Deputy Director in Informatics Research Centre, Henley Business School, University of Reading. She is a Fellow of British Computer Society, Charted Institute of IT. She is the PI and Director of EPSRC Future Blood Testing for Inclusive Monitoring and Personalized Analytics Network+. She is the academic lead of a large collaborative project of Improving the Quality of Healthcare through an Integrated Clinical Pathway Management Approach and Cloud based Digital Data Integration Platform, which was awarded ESRC O2RB Excellence in Impact Award in 2018 for her research impact on healthcare quality improvement. She has been PI and academic lead on projects funded by ESRC, EPSRC, The Health Foundation, NHS and companies, working on data-driven decision support systems that use real-world data (under privacy preserving framework) from multiple sources including Electronic Patient Record in acute, community hospital and primary care settings, remote health monitoring and patient reported outcomes to develop novel technologies (including AI based methods) to support clinical and operational decision makings in patient pathway. She got her PhD degree in informatics at Beijing Institute of Technology. She has published more than 60 papers in international journals and conferences such as European Journal of Information Systems, Journal of Information Technology, Computers in Industry, Information Systems Frontiers, Expert System with Applications.

Anthony Bradlow is a Consultant Rheumatologist recently retired from clinical practice at the Royal Berkshire Hospital, Reading. In addition to his 48 years as a clinician he was Head of the Oxford Deanery School of Postgraduate Medical Specialities for several years, as well as undertaking senior roles in clinical management and medical appraisal. He continues to undertake teaching, mentoring and appraisal as well as his recent research in the use of machine learning in rheumatology. He obtained his MB ChB and MD degrees from the University of Cape Town, South Africa and is a Fellow of the Royal College of Physicians of London.

Eghosa Bazuaye is currently the Associate Director of Informatics at the Royal Berkshire Hospital (RBH). He has built a successful track record of informatics strategy development, implementation and innovation for NHS hospitals. He is RBH informatics lead for the regional integrated care system (ICS) population health analytics programme. He leads the development of RBH's infrastructure, data marts and information governance processes to support data science and research projects. He leads the development an actionable intelligence data assistant mobile application and a clinical outcome data capture application that is now being rolled out to other Trusts in England. He is winner of the Health foundation – advancing applied analytics in health care grant 2020; speaker at the Health Strategy forum on advanced analytics and actionable intelligence 2020; winner of HSJ Value in Healthcare Awards 2014 - Improvement in Patient Information Management; shortlisted for 2015 EHI awards and 2014 NHS Innovations Award.

Antoni T.Y. Chan is Consultant Rheumatologist and Physician at the Royal Berkshire NHS Foundation Trust in Reading, United Kingdom. He did his postgraduate training at the University of Oxford where he obtained his PhD in immunology. He is also Associate Medical Director in his hospital. He has led the Outpatient Transformation in the Berkshire West Integrated Care Partnership with a strong digital focus. He successfully implemented the use of remote monitoring of blood tests across multi-specialties, collection of electronic patient reported outcome measures (e-PROM), patient management systems and bed view through data analytics from the electronic patient record, virtual clinics, telemedicine and virtual ward in his hospital and the wider healthcare system. The digital foundations started in 2017 and proved to be extremely valuable for digital readiness during the COVID-19 pandemic in 2020. His work is acknowledged by NHS England in the recently published Digital Playbook. He is a Visiting Fellow at the Henley Business School, University of Reading and a member of the Advisory Board for Blood+ Network.