Early detection of inflammatory arthritis to improve referrals using multimodal machine learning from blood testing, semi-structured and unstructured patient records

Early detection of inflammatory arthritis to improve referrals using multimodal machine learning from blood testing, semi-structured and unstructured patient records

Bing Wang
Informatics Research Centre,
Henley Business School,
University of Reading, UK
bing.wang@pgr.reading.ac.uk

Weizi Li
Informatics Research Centre,
Henley Business School,
University of Reading, UK
weizi.li@henley.ac.uk

Anthony Bradlow
Rheumatology Department,
Royal Berkshire NHS Foundation Trust,
UK
anthony.bradlow@royalberkshire.nhs.uk

Antoni T.Y. Chan
Rheumatology Department,
Royal Berkshire NHS Foundation Trust, UK
antoni.chan@royalberkshire.nhs.uk

Eghosa Bazuaye
Informatics Department,
Royal Berkshire NHS Foundation Trust, UK
eghosa.bazuaye@royalberkshire.nhs.uk

Abstract

Early detection of inflammatory arthritis (IA) is critical to efficient and accurate hospital referral triage for timely treatment and preventing the deterioration of the IA disease course, especially under limited healthcare resources. The manual assessment process is the most common approach in practice for the early detection of IA, but it is extremely labor-intensive and inefficient. A large amount of clinical information needs to be assessed for every referral from General Practice (GP) to the hospitals. Machine learning shows great potential in automating repetitive assessment tasks and providing decision support for the early detection of IA. However, most machine learning-based methods for IA detection rely on blood testing results. But in practice, blood testing data is not always available at the point of referrals, so we need methods to leverage multimodal data such as semi-structured and unstructured data for early detection of IA. In this research, we present an ensemble learning-based method using multimodal data to assist decision-making in the early detection of IA. Experimental results show the precision, recall, F1-Score, accuracy, and G-Mean of 0.89, 0.85, 0.86, 0.85, and 0.88. To the best of our knowledge, our study is the first attempt to utilize multimodal data to support the early detection of IA from GP referrals.

Keywords: Early detection of inflammatory arthritis; Ensemble learning; Multimodal data.

1. Introduction

Inflammatory arthritis (IA) and non-inflammatory conditions (NIC) are the two subdivisions of rheumatic musculoskeletal diseases (RMDs). RMDs are a group of conditions that affect the bones, joints, muscles and spine, and can cause severe long-term pain and physical disability (Van Der Heijde et al., 2018). RMDs constitute a major health problem in the general adult population due to their high prevalence and their association with significant disability, days lost at work and mortality (Government, 2022). RMDs are a common cause of long-term disability with over 20 million people in the UK (around a third of the population) living with an MSK condition, such as arthritis and low back pain (VA, 2021) and represent 46-54% of all persons with activity limitation. Approximately 1.71 billion people have RMDs conditions worldwide (WHO, 2021). In the UK, it has been estimated that rheumatoid arthritis alone costs the National Health Service (NHS) around £560 million per year and that additional costs to the economy of sick leave and work-related disability total £1.8 billion per year (Office, 2009).

Early detection, effective treatment and management of RMDs can improve the likelihood that they live in good health, and remain independent and connected to their community. To achieve this early detection and differentiation of inflammatory arthritis and non-inflammatory conditions is an essential step for the efficient and accurate referral in rheumatology from the primary care general practice (GP) to the secondary care hospitals (Wang et al., 2023). Currently, hospital referral triage in the UK dominantly relies on the manual assessment which has

URI: https://hdl.handle.net/10125/106796
978-0-9981331-7-1

to be fitted into the clinician's time and adds compound pressure on referral to treatment. In practice, patients' referral data involve various modalities, including the GP referral letter, clinical information summary (CIS), and sometimes blood test results. The manual referral assessment is labor-intensive and inefficient. Furthermore, with workforce modernization in primary care, referrals are also received from non-GP referrers, so condition descriptions in referral data can be variable in content and information. Machine learning shows great potential in the early detection and differentiation of inflammatory arthritis and non-inflammatory conditions. However, current research mainly depends on structured blood test results (Forrest et al., 2023) or imaging data (Gutiérrez-Martínez et al., 2020; Mackie et al., 2015; Ojha, Anand, & Kanisha, 2023; Shin et al., 2021) which are normally not available at the point of referrals, for example, Matsuo developed a machine learning-based method to predict the relapse of rheumatoid arthritis patients using both ultrasound and blood test data (Matsuo et al., 2022). However, these methods aren't applied for the early detection of referral improvement in practice. The prior study only includes the structured data from blood test results and the unstructured data from GP referral letters (Wang et al., 2023). But the semi-structured clinical information summary data is also useful to improve the performance of the model due to the patient's medical history included.

In this paper, we propose a multimodal machine learning method for early detection and differentiation of inflammatory arthritis and non-inflammatory conditions from blood testing, semi-structured and unstructured patient records. As far as we are concerned, there are no prior studies that consider multimodal data from the primary care GP to identify inflammatory arthritis at the point of referral. Specifically, this research makes several noteworthy contributions to the current early detection of inflammatory arthritis.

- We first introduce the clinical information summary (CIS) data into model building.
- We develop a Bidirectional Encoder Representations from Transformers (BERT) based model to process the semi-structured clinical information summary (CIS) data.
- We develop an ensemble learning-based model on the basis of the unstructured GP referral letters, structured blood test results, as well as the semi-structured clinical information summary to identify inflammatory arthritis at the point of referrals.

2. Literature Review

Early detection of IA is critical but challenging as IA can present with non-specific symptoms and there is no single current marker that is diagnostically definitive (McAllister et al., 2017).

Furthermore, the fast hospital referral triage process also requires accurate and timely disease detection from the multimodal referral data, which includes structured blood testing, semi-structured clinical information summary, and unstructured GP referral letters. However, often-large amounts of clinical information accompanying each referral have to be read and assessed by a specialist clinician to determine the appropriate care pathway for the referred patient. Processing of referrals by clinicians is labor-intensive and time-consuming, taking an average of 4.7 weeks for rheumatoid arthritis in one UK NHS hospital (Stack et al., 2019). This delay adds to existing delays to treatment due to pressure on hospital capacity. NICE Quality Standard 33 requires triage assessment by rheumatology within three weeks of referral (Nice, 2020) but the National Early Inflammatory Arthritis Audit's Second Report (2020) confirms that fewer than half of hospitals achieve target times (NRAS, 2021). Longer waits also cause significant distress for patients, who are often in considerable pain, which in turn can lead to higher costs overall because patients need intensive treatment with high-cost drugs for longer periods (NRAS, 2021).

Various tools and scales are developed for improving the efficiency of this process on the basis of the experts' clinical experience, including some general scales like the United Kingdom Birmingham Symptom specific Obstetric Triage System (BSOTS) (Kenyon et al., 2017), Manchester Triage Scale (MTS) (Mackway-Jones, Marsden, & Windle, 2013), domain-specific scales like the United States Maternal Fetal Triage Index (MFTI) (Ruhl et al., 2015), as well as some calculator-like tools such as Early Warning Systems (EWS) (O'Neill et al., 2021), Modified Early Warning Systems (MEWS) (Subbe et al., 2001), and National Early Warning Systems (NEWS) (Smith et al., 2013). However, these are still manual methods that suffer from low efficiency.

Machine learning shows great potential to improve hospital referral triage. For instance, the mortality prediction (Joseph et al., 2020), hospitalization admission predictions (Kwon et al., 2021), and severity grades classification (Zmiri, Shahar, & Taieb-Maimon, 2012). However, most of these methods aim at the hospital emergency department triage other than the triage from the GPs to hospitals. The existing research on referral triage from primary care only considered the structured data of

blood test results and the unstructured data of GP referral letters (Wang et al., 2023). But the real-world data generated is more complicated and also includes semi-structured clinical information summary (CIS) data other than just structured blood test data and GP referral letters. The semi-structured clinical information summary data contains plenty of clinical information that is essential for identifying and differentiating IA and NIC, but there is no research taking the CIS data into the model development to improve the hospital referral triage efficiency.

The CIS data consists of a series of tables or lists, which are made of patients' different medical histories, such as prescription history, disease history, allergy history, check-up history, employment, and habit histories. These data types cannot be directly used in building models due to their complex data types even in one table, but these tables and lists often have inherent structures, which contain different semantic relationships between various elements. Table 1 is an example of a part of a patient's CIS data, which includes medication, problems, allergies, and social context components. For simplicity, only several records of each component are shown in this table.

In this research, we first propose a multimodal model for the early detection of IA by using the CIS data, blood test data, as well as GP referral letters from the patients' referral data.

Table 1. Example of a patient's clinical information summary (CIS) data, including four tables, medication, problems, allergies, and social context.

(1) Medication table

Drug	Dosage	Quantity	Last Issued On	Type
Fluoxetine 10mg tablets	One To Be Taken Each Day (to make dose up to 30)	30 tablets	11-Dec-18	Acute
Oxactin 20mg capsules (Discovery Pharmaceuticals)	One To Be Taken Each Day	56 capsules	10-Dec-18	Repeat

(2) Problems table

Date	Problem	Associated Text	Date Ended	Type
12-Mar-10	Low mood	still hasles with sister		Active
14-Sep-09	Weight symptom	also hot flushes		Active

(3) Allergies table

Date	Description	Associated Text
25-Mar-13	Adverse reaction to Fluoxetine Hydrochloride	suspected fluoxetine induced rash

(4) Social Context table

Category	Date	Description	Value	Units	Range
Smoking	01-Dec-14	Never smoked tobacco			
Alcohol	19-Aug-15	Alcohol consumption	2	U/week	

3. Methodology

As illustrated in Figure 1, the proposed model consists of four components, including two separate BERT-based classification models to classify the patients with either GP referral letters or CIS data, a tree-based model to identify the patients with blood test data, and an ensemble learning-based model to generate the final prediction. Specifically, two BERT-based models are separately fine-tuned for GP referral letters and CIS data, and a gradient boosting machine (GBM) classifier is trained for blood testing data. Furthermore, an ensemble learning module is used to

get the final prediction based on the predictions from the single classifiers.

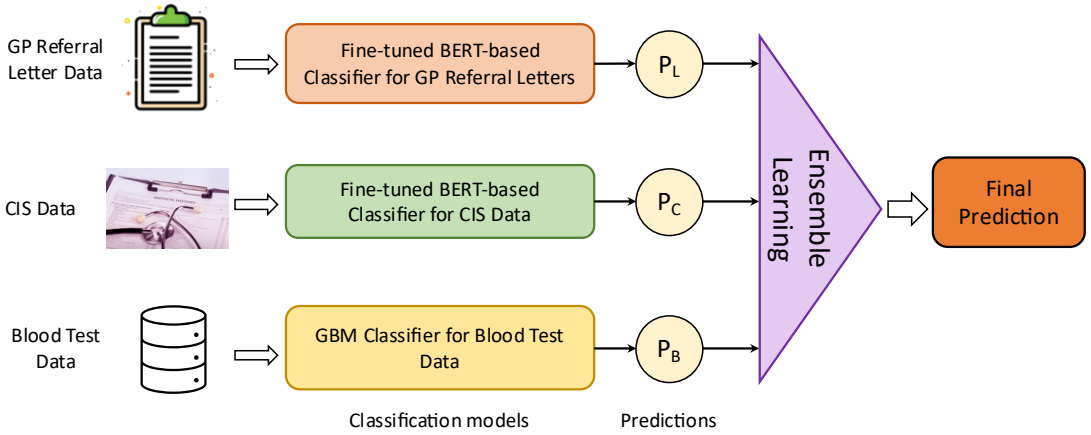

Figure 1. Multimodal machine learning model for the early detection of inflammatory arthritis.

3.1. BERT-based Classification Model

As the state-of-the-art pre-trained language model, BERT is developed on a multi-layer bidirectional transformer encoder architecture, which is based on a self-attention mechanism (Devlin et al., 2019). Generally, there are two steps involved in the development of a BERT-based classification model, including the pretraining procedure and the fine-tuning procedure. For the pretraining BERT procedure, two unsupervised tasks are used, including the masked language model (MLM) and next sentence prediction (NSP), to train a BERT model on a relatively large corpus, such as BooksCorpus, and Wikipedia. For the fine-tuning BERT procedure, a pre-trained BERT-base or BERT-large model is used like a feature extractor to fine-tune the downstream tasks such as text classification and question answering, on a relatively small and labeled dataset.

In this research, two BERT-based classification models are separately fine-tuned based on the pre-trained BERT-base model by using unstructured GP referral letters and semi-structured clinical information summary data. We use the first token embedding as the text representation of the input sentence and append a fully connected layer for further feature extraction and a SoftMax layer for the classification purpose. Considering the overfitting problem on the small dataset, the parameters of the lower layers of the BERT model are frozen and only the higher layers' parameters would be updated during the training process.

3.2. Gradient Boosting Machine (GBM) Model

We used the gradient boosting machine (GBM) model to classify the patients having IA or NIC based on the blood test data other than the other machine learning model due to its superior performance compared to other methods (Rufo et al., 2021; Rybarczyk & Zalakeviciute, 2021). First, it will be non-essential to perform the missingness imputation and features scaling because the GBM can handle the original data with missing values and can select features internally. Furthermore, GBM could provide the feature importance internally, which could make the predictions of the machine learning model reliable and convincible for health care workers because most of the machine learning models are black-box models, and there are no underlying logics provided for the doctors.

In this research, LightGBM is chosen to implement the classification model due to its outstanding features compared with other tree-based boosting frameworks (Rufo et al., 2021). First, the leaf-wise algorithm of the LightGBM can reduce loss and thus improve the accuracy in comparison with other boosting algorithms, such as depth-wise or level-wise algorithms. Furthermore, LightGBM can provide faster training speed, higher efficiency, distributed training, as well as GPU support.

Although GBM is a nonparametric approach and shows great performance in practice, there are various hyperparameters that would affect the final accuracy. To search for the optimal hyperparameters of the machine learning model, the Ray Tune Python

package is chosen to search for the optimal hyperparameters of the GBM classifier (Moritz et al., 2018).

3.3. Ensemble Learning Methods

Ensemble learning is a commonly used method to improve the model's accuracy, and shows great potential in boosting the generalization capabilities of the model. Modern ensemble learning research mainly focuses on feature fusion to investigate the complex supplementary and complementary relationships between different data modalities, which involves creating new feature representation from multi-source data or models, such as concatenation, element-wise addition and multiplication, linear and nonlinear transformations. The traditional ensemble learning methods aim at combining the predictions of different single classifiers to get a better performance compared with the single classifiers, which includes bagging, boosting, and stacking. In this research, the voting strategy is applied to fuse the different classifiers' predictions, and soft voting (SV) and hard voting with the threshold (HVT) are compared in this research.

$$\widehat{P_{HVT}} = mode_t(P_L, P_C, P_B), t \in (1, 2, 3)$$

$$\widehat{P_{SV}} = \arg\max_{i=(NIC,IA)} \sum_{j=(L,C,B)} p_{i,j}$$

Where $\widehat{P_{HVT}}$ and $\widehat{P_{SV}}$ represent the HVT and SV voting methods separately, and for hard voting with the threshold (HVT), $mode_t$ function means the hard voting with the threshold of t.

3.4. Performance Metrics

In this study, different kinds of metrics are used to measure the performance of the model, including Accuracy, Precision, Recall, and F1-Score, which are calculated by the following equations.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F_1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$Sensitivity = \frac{TP}{TP + FN}$$

where TP, FP, FN, and TN represent true positive (TP), false positive (FP), false negative (FN), and true negative (TN) from the confusion matrix, as shown in Table 2.

Table 2. Confusion matrix.

	Real NIC	Real IA
Predict NIC	TN	FN
Predict IA	FP	TP

3.5. Hardware and Software

The BERT-based models are fine-tuned by using the NVIDIA V100 graphics processing unit (GPU) as well as Intel CPU on the research computation cluster, and the LightGBM model is trained on the laptop computer with specifications of Intel Core i7 CPU and 16GB RAM. We used different Python packages to develop the programs, mainly including Pytorch (Version 1.12.1), Transformers (Version 4.23.1), Scikit-learn (Version 1.2.2), Ray Tune (Version 2.4.0), and LightGBM (Version 3.3.3).

4. Experiment

4.1. Datasets

We employed a dataset obtained from the Rheumatology Department of a large secondary care hospital in the United Kingdom. All patients' referral data were referred from GPs to the secondary care hospital between February 2018 and July 2021. We choose the referral triage data of inflammatory and non-inflammatory conditions as the application case. The reason is that inflammatory arthritis (IA) is one of the most common chronic diseases and can often impose a severe burden on individuals and society (Van Der Heijde et al., 2018).

All data were anonymized according to regulations of data protection and information governance required by the hospital's compliance department. Specifically, the dataset is split into four parts, including an ensemble dataset having three data modalities (GP referral letter, CIS data, and blood test data), and three separate datasets having single data modality. The ensemble dataset will be used for ensemble learning, and the three unimodal datasets will be used to train two BERT-based models and a tree-based model separately. A detailed summary of the datasets could be found in Table 3. Specifically, the ensemble dataset includes 165 patients with 39 inflammatory arthritis and 126 non-inflammatory conditions. For the unimodal datasets, we have 1199, 740, and 1028 of the GP letter dataset, CIS dataset, and blood test dataset. The GP letter data is the general natural language description of the patient's physical

conditions when they came to the primary care GP clinics, which mainly consists of GP's consideration and simple physical check-ups. The CIS data is the textual description of the patient's previous medical history, such as the prescription history and disease history. The blood test data include some simple lab test results as well as demographical information like age and gender.

Table 3. Summary of the dataset.

Dataset Types	NIC	IA	Total
GP letter dataset	837	362	1199
CIS dataset	540	200	740
Blood test dataset	236	792	1028
Ensemble dataset	126	39	165

Note: Non-inflammatory condition (NIC), Inflammatory arthritis (IA).

4.2. Data Preprocessing

Different data preprocessing strategies were applied for the various data modalities. For the unstructured GP referral letters, GP referral letters without clinical information will be excluded from our research, and then some special characters will be removed to format the text data, such as the line break, and additional whitespace. For the semi-structured clinical information summary, there are four sections (medications, problems, allergies, and social context) and each of them is a semi-structured form. The contents of different tables are concatenated after sorting by time, which will be used to fine-tune a BERT-based classification model. For the structured blood test results, previous data within one year before the point of the triage are used to impute a portion of the missing values, and then some invalid values without any clinical meaning are dropped. Finally, there are totally 29 features used for model development, including patients' demographic information, haematology (routine), blood biochemistry (routine) and immunology.

4.3. Unimodal Method Performance

We tested three single classifiers trained on different sub-datasets, including the GP referral letter dataset, CIS dataset, and blood test results dataset, and the detailed performances are shown in Table 4.

Overall, we applied the same data partition tactics for three datasets. Specifically, 80% of the data are used for training and 20% for the testing purpose, and then 20% of the training set will be used as the evaluation dataset to tune the threshold, and the rest of the 80% of the training set will be used for training models. Notably, the lower layers of the fine-tuned BERT architecture for the GP referral letter dataset and the CIS dataset are frozen and only the top 5 layers of the original BERT model are fine-tuned.

Clearly, for single classifiers directly trained on the separate datasets, the GBM model using blood test results achieves the best performance compared with the fine-tuned BERT classifiers using either GP referral letters or the CIS data, with the G-Mean and AUC values of 0.77 and 0.84.

Table 4. Performance of the unimodal methods.

Models	**Precision**	**Recall**	**F1-Score**	**Accuracy**	**AUC**	**G-Mean**
Fine-tuned BERT for GP Referral Letter	0.75	0.69	0.70	0.69	0.80	0.70
Fine-tuned BERT for CIS	0.73	0.60	0.62	0.60	0.70	0.64
GBM model for Blood Test Results	0.84	0.83	0.83	0.83	0.84	0.77

4.4. Multimodal Method Performance

Overall, we compared two kinds of ensemble learning methods, including hard voting and soft voting. Slightly different from the general hard voting method, we tested hard voting with different thresholds, which means how many single classifiers correctly predict the patients having IA. From Table 5, the hard voting with a threshold of one shows the best G-Mean of 0.88, which is significantly better than other methods. Specifically, our ensemble model achieves the weighted precision, recall, F1-score, and accuracy of 0.89, 0.85, 0.86, and 0.85.

Table 5. Comparison results of various ensemble approaches.

Methods	Precision	Recall	F1-Score	Accuracy	G-Mean
Soft Voting (SV)	0.81	0.82	0.79	0.82	0.57
Hard Voting with Threshold (HVT@1)	0.89	0.85	0.86	0.85	0.88
Hard Voting with Threshold (HVT@2)	0.81	0.82	0.79	0.82	0.58
Hard Voting with Threshold (HVT@3)	0.83	0.79	0.71	0.79	0.32

Note: Precision, Recall, and F1-Score in the above table are weighted performance metrics. HVT@1, HVT@2, and HVT@3 mean hard voting with different thresholds, including one, two, and three.

5. Discussion and Future Work

Inflammatory arthritis (IA) is a type of rheumatic musculoskeletal diseases (RMDs) characterized by joint inflammation. It is estimated that 1.71 billion people suffer from RMDs condition worldwide. The early detection of inflammatory arthritis during the hospital referral triage from the GP referrals is significantly essential to prevent the disease course from worsening and help patients acquire timely treatments. However, it is challenging to build a high-performing decision support model using the referral data from GPs due to the various data modalities challenges, as well as the data incompleteness challenges involved in the hospital referrals.

This study makes several contributions to addressing these challenges. First, we developed a multimodal machine learning method using patients' GP referral letters, clinical information summary (CIS), and blood test data for the early detection of inflammatory arthritis at the point of hospital referral triage from primary care GPs. Second, our research proposes a BERT-based model to analyze the patient's CIS data. The experiments show that our multimodal method achieves the G-Mean, precision, recall, F1-score, and accuracy values of 0.88, 0.89, 0.85, 0.86, and 0.85.

Our research lays the groundwork for future research on several aspects. First, we will investigate the data missingness challenges to improve the model's performance and generalization capabilities. This includes not only the unimodal data missingness, for example, the missing values for some specific attributes in the blood test results, but also the missingness of the data modalities, for instance, some patients miss some specific data modalities.

Second, a more robust and advanced attention-based multimodal fusion architecture will be developed to improve the performance of the model. Besides, explainable models will be developed to explain the predictions of the proposed mode thus that it will help clinicians back-trace disease predictions (e.g., salient factors) for transparent triaging recommendations.

Third, we will launch live hospital validation and further human versus computer trials once we obtained superior experimental results based on a larger dataset.

Furthermore, in the long term, we hope to extend this method to wider hospitals and departments that are under pressure of manual referral triage and have a large number of referrals. Moreover, we anticipate supplementing data from other modalities to further explore the model performance and applications. For example, the electronic patient-reported outcome measures (ePROMs) data, which is time series and can be used to build the model so that disease outcome forecasting and personalized recommendations (e.g., follow-ups) can be provided.

With the digital transformations in hospitals worldwide, our method shows great potential in saving clinicians' time, improving the efficiency of the referral management, ensuring that patients are diagnosed rapidly, and improving patient outcomes. For example, it currently takes clinicians about 15 hours a week (=3.75 clinics) screening GP referrals, and decide on follow-up patients at our collaborating secondary care hospital in the UK. Time will be saved for clinicians for direct patient care through automatic decision support.

6. Acknowledgement

This work is funded by Health Innovation Partnership between Royal Berkshire NHS Foundation Trust and University of Reading; EPSRC Future Blood Testing Network for Inclusive Monitoring and Personalized Analytics Network+ (EP/W000652/1); National Institute of Health and Care Research (NIHR), invention for innovation (i4i) FAST (Funding at the Speed of Innovation) (NIHR205854); and Engineering and Physical

Sciences Research Council (EPSRC) (EP/Y019393/1); National Natural Science Foundation of China (No.72074023). We appreciate the High Performance Computation (HPC) services provided by the Reading Academic Computing Cluster (RACC).

7. References

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, jun). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, Minnesota.

Forrest, I. S., Petrazzini, B. O., Duffy, Á., Park, J. K., O'Neal, A. J., Jordan, D. M., Rocheleau, G., Nadkarni, G. N., Cho, J. H., & Blazer, A. D. (2023). A machine learning model identifies patients in need of autoimmune disease testing using electronic health records. *Nature Communications*, *14*(1), 2385.

Government. (2022). *Musculoskeletal health: applying All Our Health*. https://www.gov.uk/government/publications/musculoskeletal-health-applying-all-our-health/musculoskeletal-health-applying-all-our-health

Gutiérrez-Martínez, J., Pineda, C., Sandoval, H., & Bernal-González, A. (2020). Computer-aided diagnosis in rheumatic diseases using ultrasound: an overview. *Clinical Rheumatology*, *39*(4), 993-1005.

Joseph, J. W., Leventhal, E. L., Grossestreuer, A. V., Wong, M. L., Joseph, L. J., Nathanson, L. A., Donnino, M. W., Elhadad, N., & Sanchez, L. D. (2020). Deep-learning approaches to identify critically Ill patients at emergency department triage using limited information. *Journal of the American College of Emergency Physicians Open*, *1*(5), 773-781.

Kenyon, S., Hewison, A., Dann, S.-A., Easterbrook, J., Hamilton-Giachritsis, C., Beckmann, A., & Johns, N. (2017). The design and implementation of an obstetric triage system for unscheduled pregnancy related attendances: a mixed methods evaluation. *BMC pregnancy and childbirth*, *17*(1), 1-10.

Kwon, J.-m., Jeon, K.-H., Lee, M., Kim, K.-H., Park, J., & Oh, B.-H. (2021). Deep learning algorithm to predict need for critical care in pediatric emergency departments. *Pediatric emergency care*, *37*(12), e988-e994.

Mackie, S. L., Koduri, G., Hill, C. L., Wakefield, R. J., Hutchings, A., Loy, C., Dasgupta, B., & Wyatt, J. C. (2015). Accuracy of musculoskeletal imaging for the diagnosis of polymyalgia rheumatica: systematic review. *RMD open*, *1*(1), e000100.

Mackway-Jones, K., Marsden, J., & Windle, J. (2013). *Emergency triage: Manchester triage group*. John Wiley & Sons.

Matsuo, H., Kamada, M., Imamura, A., Shimizu, M., Inagaki, M., Tsuji, Y., Hashimoto, M., Tanaka, M., Ito, H., & Fujii, Y. (2022). Machine learning-based prediction of relapse in rheumatoid arthritis patients using data on ultrasound examination and blood test. *Scientific Reports*, *12*(1), 7224.

McAllister, K., Goodson, N., Warburton, L., & Rogers, G. (2017). Spondyloarthritis: diagnosis and management: summary of NICE guidance. *BMJ*, *356*, j839. https://doi.org/10.1136/bmj.j839

Moritz, P., Nishihara, R., Wang, S., Tumanov, A., Liaw, R., Liang, E., Elibol, M., Yang, Z., Paul, W., & Jordan, M. I. (2018). Ray: A distributed framework for emerging {AI} applications. 13th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 18),

Nice. (2020). *Rheumatoid arthritis in over 16s*.

NRAS. (2021). *The National Early Inflammatory Arthritis Audit (NEIAA)*.

O'Neill, S., Clyne, B., Bell, M., Casey, A., Leen, B., Smith, S. M., Ryan, M., & O'Neill, M. (2021). 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 emergency medicine*, *21*(1), 1-19.

Office, N. A. (2009). Services for people with rheumatoid arthritis. *National Audit Office*, *823*, 1-37.

Ojha, S., Anand, S., & Kanisha, B. (2023). Prediction of Rheumatoid Arthritis using Deep Learning Techniques. 2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC),

Rufo, D. D., Debelee, T. G., Ibenthal, A., & Negera, W. G. (2021). Diagnosis of diabetes mellitus using gradient boosting machine (LightGBM). *Diagnostics*, *11*(9), 1714.

Ruhl, C., Scheich, B., Onokpise, B., & Bingham, D. (2015). Content validity testing of the maternal fetal triage index. *Journal of Obstetric, Gynecologic & Neonatal Nursing*, *44*(6), 701-709.

Rybarczyk, Y., & Zalakeviciute, R. (2021). Assessing the COVID-19 impact on air quality: A machine learning approach. *Geophysical Research Letters*, *48*(4), e2020GL091202.

Shin, Y., Yang, J., Lee, Y. H., & Kim, S. (2021). Artificial intelligence in musculoskeletal ultrasound imaging. *Ultrasonography*, *40*(1), 30.

Smith, G. B., Prytherch, D. R., Meredith, P., Schmidt, P. E., & Featherstone, P. I. (2013). 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), 465-470.

Stack, R. J., Nightingale, P., Jinks, C., Shaw, K., Herron-Marx, S., Horne, R., Deighton, C., Kiely, P., Mallen, C., & Raza, K. (2019). Delays between the onset of symptoms and first rheumatology consultation in patients with rheumatoid arthritis in the UK: an observational study. *BMJ open*, *9*(3), e024361. https://doi.org/10.1136/bmjopen-2018-024361

Subbe, C. P., Kruger, M., Rutherford, P., & Gemmel, L. (2001). Validation of a modified Early Warning Score in medical admissions. *Qjm*, *94*(10), 521-526.

VA. (2021). *The State of Musculoskeletal Health*. https://www.versusarthritis.org/about-arthritis/data-and-statistics/the-state-of-musculoskeletal-health/

Van Der Heijde, D., Daikh, D. I., Betteridge, N., Burmester, G. R., Hassett, A. L., Matteson, E. L., Van Vollenhoven, R., & Lakhanpal, S. (2018). Common language description of the term rheumatic and musculoskeletal diseases (RMDs) for use in communication with the lay public, healthcare providers and other stakeholders endorsed by the European League Against Rheumatism (EULAR) and the American College of Rheumatology (ACR). *Annals of the rheumatic diseases*, *77*(6), 829-832.

Wang, B., Li, W., Bradlow, A., Bazuaye, E., & Chan, A. T. (2023). Improving triaging from primary care into secondary care using heterogeneous data-driven hybrid machine learning. *Decision Support Systems*, *166*, 113899.

WHO. (2021). *Musculoskeletal conditions*. https://www.who.int/news-room/fact-sheets/detail/musculoskeletal-conditions

Zmiri, D., Shahar, Y., & Taieb-Maimon, M. (2012). Classification of patients by severity grades during triage in the emergency department using data mining methods. *Journal of evaluation in clinical practice*, *18*(2), 378-388.