Multi-stage multimodal fusion network with language models and uncertainty evaluation for early risk stratification in rheumatic and musculoskeletal diseases

Information Fusion 120 (2025) 103068

Contents lists available at ScienceDirect

Information Fusion

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

Multi-stage multimodal fusion network with language models and uncertainty evaluation for early risk stratification in rheumatic and musculoskeletal diseases

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

[a] Informatics Research Centre, University of Reading, Reading, RG6 6UD, UK
[b] Rheumatology Department, Royal Berkshire NHS Foundation Trust, Reading, RG1 5AN, UK
[c] Nuffield Department of Clinical Medicine, University of Oxford, Oxford, OX3 7BN, UK
[d] Informatics Department, Royal Berkshire NHS Foundation Trust, Reading, RG1 5AN, UK

ARTICLE INFO

Keywords:
Multi-stage multimodal fusion
language models
Uncertainty quantification
Conformal prediction
Early risk stratification of RMDs

ABSTRACT

Precise risk stratification of rheumatic musculoskeletal diseases (RMDs) is crucial for ensuring patients get right referrals and treatments quickly. However, it is challenging due to the non-specific symptoms and the lack of the diagnostically definitive single biomarker. The real-world referral data present several challenges such as the free format texts and incomplete data challenges, which introduces further modeling complexity, and makes uncertainty quantification crucial for ensuring reliable predictions and outcomes. To solve these challenges, we developed a multi-stage multimodal fusion network with conformal prediction method that can accurately risk stratify RMDs at the point of referrals, quantify the uncertainty and flag unreliable predictions for physician's interventions. The proposed models were trained and evaluated using referral data from 128 General Practices (GPs) in the UK, which include patients who visited and were referred by GPs with suspected inflammatory conditions in RMDs between February 2018 and January 2024. Our model achieved 0.73 accuracy, 0.79 AUC, and 0.75 G-Mean to differentiate inflammatory conditions (IC) and non-inflammatory conditions (NIC) using patients' presenting condition description (PCD) and medical history (MH) data, and 0.90 accuracy, 0.92 AUC, and 0.89 G-Mean using patients' PCD, MH and additional blood test data (BTD). Furthermore, conformal prediction-based method has been developed to evaluate prediction uncertainty and can further identify 75.71 % unreliable predictions for patients with PCD and MH data, and 66.67 % unreliable predictions for patients with additional BTD data, which could be given a second-round examination by GP/secondary care clinicians for patient safety. The findings of this study suggest that language models with multi-stage multimodal fusion and uncertainty evaluation can risk stratify RMDs accurately using data available at the point of referral in the real world. Therefore, it is possible to be used by GPs and clinicians to help patients get the right treatment faster, demonstrating practical potential to improve RMDs referrals in the real world.

1. Introduction

Rheumatic and musculoskeletal diseases (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. RMDs are a common cause of long-term disability and over 20 million people in the UK (around a third of the population) [1,2] live with RMDs [3]. Approximately 1.71 billion people have RMD conditions worldwide [4]. RMD conditions are the leading contributor to disability and unemployment worldwide [5] and are the most common medical causes of long-term absence from work, accounting for more than half of all sickness [6–10].

Inflammatory conditions (IC, mainly but not exclusively inflammatory arthritis) and non-inflammatory conditions (NIC) are the two major subdivisions of RMDs, each with very different treatment and management pathways (e.g., disease-modifying drugs for inflammatory conditions e.g., rheumatoid arthritis; surgeries such as joint replacements for non-inflammatory conditions e.g., osteoarthritis). Accurate early

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

https://doi.org/10.1016/j.inffus.2025.103068
Received 7 October 2024; Received in revised form 22 February 2025; Accepted 26 February 2025
Available online 1 March 2025

detection and differentiation of IC and NIC are critical for patients to be referred to the right specialists and receive the right treatment rapidly. However, early detection is challenging because IC and NIC often present with non-specific symptoms and there is currently no diagnostically definitive single biomarker to detect inflammatory arthritis [11].

Machine learning (ML) based risk stratification has already shown superior performance to support disease diagnosis without significant labor cost and enhanced accuracy [12,13], but evidence confirms there are significant gaps in the field for ML-based early detection [14]. Current ML research in RMD focuses on deep learning-based imaging analysis to support diagnosis of osteoarthritis and rheumatoid arthritis using CT, MRI, X ray and ultrasound data [15–19], while other studies to predict RMDs like rheumatoid arthritis using different clinical data [20], such as blood test results [21,22], and genetic data [23,24]. However, imaging examinations and specialized blood testing are usually conducted after referral to secondary care. These tests are not conducted frequently on the patient's first visit to the primary care practitioner. Therefore, existing ML studies still rely on the data obtained through advanced testing and imaging and thus do not fit for early detection in practice when the symptoms first appear, and our study is the first of this kind to use multimodal data from the primary care practitioner to differentiate the early IC and NIC. Furthermore, real-world early symptom data for early detection is always unstructured text and in heterogeneous formats, such as the unstructured presenting conditions description in GP referral letters and semi-structured medical history in clinical information summary [25].

Although recent studies about large language models have shown promising performance in disease diagnosis by mimicking common clinical reasoning processes of physicians, they were only applicable with predesigned and structured symptom checkers [26] and question-and-answer datasets rather than real-world clinical data in heterogenous formats [27,28]. In addition, multi-modal data fusion, such as the attention-based feature fusion network, has proven to be an effective method in medical data analysis in terms of its promising capability to extract complementary information among various data modalities [29,30]. Attention-based fusion strategies demonstrate effectiveness primarily on complete datasets. Some multimodal machine learning models have been developed to detect some early RMDs. For example, a joint multi-modal learning method has been developed for the classification of the grade of early-stage knee osteoarthritis disease [31], an explainable multimodal learning framework has been proposed to enhance osteoporosis detection [32], and a rheumatoid arthritis knowledge guided system has been developed to score the RA activity from multimodal ultrasound images [33]. However, these studies are not suited for our research due to several limitations: (1) These research focus on the diagnosis at the secondary care hospitals instead of the primary care practitioners; (2) These studies require complete datasets for modeling, such as the comprehensive clinical and imaging data which are not available at the referral stage and (3) These studies do not quantify the uncertainty of the model predictions, and the contribution of various data modalities. Furthermore, in real-world applications, the incomplete data challenges significantly hinder the implementation of these methods. This limitation is particularly evident in the early diagnosis of certain diseases like RMDs in our study, where a substantial proportion of patients present with incomplete data types such as blood test results meaning they do not have blood test results available during referrals.

In addition, for GPs and clinicians to use and trust the ML-based risk stratifications for RMD in real-world clinical practice, it is critical [34] to evaluate prediction uncertainty and detect unreliable predictions [35]. Real-world applications will inevitably expose the ML model to clinical data beyond the data upon which they were trained, either because it is unusual (e.g., a rare RMD case previously unseen for the model) or because it originates from an evolving patient population. The ability to evaluate uncertainty and detect unreliable predictions is key in ensuring patient safety of the ML model in real-world RMD referrals. However, there is limited research evaluating prediction uncertainty from unstructured data with different modalities, and evaluating the prediction uncertainty in these methods is challenging due to the complexity and heterogeneity of the data sources, as well as the interactions between modalities that can introduce additional uncertainty.

To address these challenges, we aimed to develop a language model and multimodal fusion-based method to differentiate IC and NIC patients early using heterogenous referral data from primary care general practitioners. We further developed the conformal predictors to detect unreliable predictions by analyzing the uncertainty of the model's predictions. Our methodology will be used as a decision support system to show the likelihood of each patient having IC or NIC, with prediction reliability assessment to ensure patient safety and clinical utility [36]. If the system is successfully adopted in the real world, the ability to detect and differentiate IC and NIC early will enable more accurate referrals and patients can be correctly treated more quickly.

Our main contributions can be summarized as follows:

- Language modelling for unstructured data in different modalities. Our research fills the gap of limited study in the early risk stratification of RMDs in primary care, and we propose a disease early detection model that encodes different textual modalities in primary care based on language models.
- Multimodal and multistage fusion to address data challenges in primary care. We employ the attention based fusion method to enhance the latent representation of the multimodal textual data embedding by integrating existing language models. The attention-based methods can dynamically extract complementary information between different feature representations, enabling the model to focus on the most relevant aspects of each modality. Furthermore, multistage fusion methods are used to incorporate incomplete data type in the multimodal fusion in real-world healthcare.
- Uncertainty evaluation of predictions from unstructured and multimodal data. We employ the conformal prediction together with the proposed models to evaluate the prediction uncertainty and guarantee the error rate is bounded by a pre-specified level. Our method will identify unreliable predictions that may pose risks or lead to incorrect clinical decisions, allowing clinicians to take corrective measures or flag them for further review. Our methods ensure patient safety in implementing multimodal methods in the real world [35].
- Explaining predictions from multimodal data. We apply the SHapley Additive exPlanations (SHAP) [37] based explanations to quantify and explain contributions to predictions from multimodal data. This will enhance the transparency and interpretability of multimodal machine learning models. Specifically, our approach can identify critical words and key indicators in unstructured, semi-structured and structured referral data that contribute most to model's predictions.
- To the best of our knowledge, our model is the first to detect RMD disease early using real-world referral data. Our method has been evaluated on a retrospective dataset that has been collected from 128 GPs in the UK.

2. Literature review

2.1. Referral of rheumatic musculoskeletal diseases in National Health Service

There were an estimated 329 million appointments in primary care in 2022 in the UK [38] and RMDs accounts for more than 20 % of GP consultations [39], which means there are an estimated 65.8 million RMD-related GP appointments a year. Due to non-specific symptoms and a lack of biomarkers for detecting and differentiating IC and NIC in RMD [11], only 40 % of suspected early IC patients referred by GPs in 2019/2020 are proved to be accurate [40]. Inaccurate referrals can lead

to longer times for patients to gain access to the right clinics, which often results in a loss of the window of opportunity for effective treatment. Patients often consult GP multiple times while awaiting their specialist review and treatment. It is estimated that one in three GP appointments are for patients waiting for hospital services [41]. This has increased GP workload significantly, e.g. there could be around 21.9 million RMD appointments that are unnecessary. It also affects recruitment and retention within the GP profession [42]. The risk stratification tool that detects and differentiates IC and NIC at the point of referrals will support GPs to refer patients accurately, reduce delays to treatment and reduce unnecessary repeating consultations.

2.2. Multimodal fusion machine learning in healthcare

Multimodal fusion is increasingly becoming a common technique in multimodal representation learning in healthcare [43], where data from various modalities are fused to enhance prediction performance [44]. The fusion process can happen at different stages of the modeling process, such as the early fusion, intermediate fusion, and late fusion [45]. Early fusion is easy to implement, and involves concatenating input modalities or features before any processing. Although this kind of methods are straightforward, these approaches may not be suitable for complex data modalities [43]. A more advanced technique, known as intermediate fusion (or joint fusion), can capture joint feature interactions like supplementary and complementary information from intermediate layers of networks between different data modalities [46]. However, these strategies require complete data from each modality which are often impossible in the real-world healthcare. Another alternative is late fusion, where separate models are trained for each modality, and the output probabilities are combined instead of combining the original data (early fusion) or learned feature representation (joint fusion). This method is simple and robust only when good marginal models could be learned from the specific modalities, but it suffers from learning the multimodal effects on data or feature level [47]. Our proposed multistage multimodal fusion methods can enhance adaptability and practical usability in real-world applications by modeling the complementary relations between different complete and incomplete data modalities.

2.3. Quantification of predictive uncertainty in healthcare

Machine learning has shown excellent performance in healthcare, leading to more accurate disease diagnoses [48]. However, the "black box" nature of most AI systems raises concerns particularly in fields like healthcare and medicine [49]. To mitigate these issues, explainable AI (XAI) has been introduced to enhance the transparency of machine learning models [49], for example, XAI has been used to enhance the transparency and interpretability in AI-driven lung disease diagnosis, which can bridge the gap between complex AI models and clinical settings [50]. Despite this, XAI falls short in providing a practical evaluation of the reliability of the models' predictions [51,52]. Consequently, uncertainty quantification has been proposed to assess the confidence in predictions made by machine learning systems, ensuring their safety and reliability in real-world healthcare applications [53].

Predictive uncertainty is a commonly utilized technique while making predictions or estimates using a model [48], and can quantify the level of confidence or reliability in the model's predictions for new or unseen data [48]. To quantify the prediction uncertainty of machine learning models, various methods have been proposed, such as Monte Carlo simulation and Bayesian inference [54,55]. However, these methods suffer from computational costs and complex modeling process. In this study, Conformal Predictions (CP) is chosen due to several advantages [35]: (1) Unlike Bayesian inference and Monte Carlo Dropout, which rely on specific model assumptions (e.g., prior distributions, model architecture) or training modifications, conformal prediction methods are assumption-free and can be applied post-hoc to any model, which enhances their versatility and ease of integration into existing workflows [56]; (2) Conformal predictions can provide valid uncertainty estimates with formal guarantees on prediction interval coverage, regardless of the underlying model, and adjustable confidence levels to reflect the clinical cost of erroneous predictions; (3) Conformal predictions are computationally efficient and lightweight, especially when applied to pre-trained models, compared with the computationally intensive and complex modeling features of the Bayesian inference and Monte Carlo Dropout; (4) The prediction intervals generated by conformal methods are straightforward to interpret, making them particularly valuable for real-world applications where decision-makers may not have expertise in probabilistic modeling, for example, conformal prediction is used for the image segmentation with predictive uncertainty borders [57]. To the best of our knowledge, no existing research has applied CP to language models and multimodal models. Thus, we are the first to incorporate CP into multi-stage multimodal machine learning models to evaluate predictive uncertainty based on unstructured text data, advancing the application of conformal prediction in healthcare research.

3. Methodology

3.1. Dataset preparation

The clinical referral data were retrospectively collected from 128 GPs in the UK from February 2018 to January 2024. Patient's referral data include presenting condition description (PCD), medical history (MH), and available blood test data (BTD). Specifically, the PCD includes the patients' symptoms, location, duration, intensity, and accompanying symptoms when they visit the general practitioners. MH is about patients' previous clinical history, including medication, problems, allergies, consultation, and social contexts. Some patients who had blood tests completed may also have BTD data ready during referrals. A detailed description of the clinical referral data can be found in the **Section S1** in **Supplementary Material**.

Our study population included 5007 patients (mean [SD] age, 62.1 [17.6]), of which 1893 patients having non-inflammatory conditions (NIC) (mean [SD] age, 59.2 [17.4] years) and 3114 patients having inflammatory conditions (IC) (mean [SD] age, 63.8 [17.5] years) (Table 1). Specifically, men and women represented 1642 (32.8 %) and

Table 1
Statistical characteristics of the datasets.

Subgroups	IC (n, %)	NIC (n, %)	Total (n, %)
Age, mean (SD)	63.8 (17.5)	59.2 (17.4)	62.1 (17.6)
Gender			
Male	1135 (36.4)	507 (26.8)	1642 (32.8)
Female	1940 (62.3)	1356 (71.6)	3296 (65.8)
Unknown	39 (1.3)	30 (1.6)	69 (1.4)
Race			
Asian	134 (4.3)	73 (3.9)	207 (4.1)
Black	26 (0.8)	20 (1.1)	46 (0.9)
Mixed	7 (0.2)	3 (0.2)	10 (0.2)
White	2427 (77.9)	1441 (76.1)	3868 (77.3)
Other Ethnicity	52 (1.7)	50 (2.6)	102 (2.0)
Not Stated	359 (11.5)	205 (10.8)	564 (11.3)
Unknown	109 (3.5)	101 (5.3)	210 (4.2)
Index of Multiple Deprivation (IMD)			
1	53 (1.7)	24 (1.3)	77 (1.5)
2	67 (2.2)	31 (1.6)	98 (2.0)
3	178 (5.7)	120 (6.3)	298 (6.0)
4	206 (6.6)	113 (6.0)	319 (6.4)
5	173 (5.6)	99 (5.2)	272 (5.4)
6	181 (5.8)	117 (6.2)	298 (6.0)
7	344 (11.0)	174 (9.2)	518 (10.3)
8	238 (7.6)	170 (9.0)	408 (8.1)
9	491 (15.8)	313 (16.5)	804 (16.1)
10	984 (31.6)	569 (30.1)	1553 (31.0)
Unknown	199 (6.4)	163 (8.6)	362 (7.2)

3296 (65.8 %) of the patients, and 69 (1.4%) patients' gender information were missing. The research population included 207 Asian patients (4.1 %), 46 Black patients (0.9 %), 10 patients (0.2 %) with mixed ethnical backgrounds, 3868 White patients (77.3 %), 102 Other Ethnicity participants (2.0 %), 564 patients (11.3 %) that the ethnical background not stated, and 210 patients (4.2 %) with ethnicity information missing. Furthermore, the patients were from areas with indices of multiple deprivation (IMD) [58] including 77 (1.5 %), 98 (2.0 %), 298 (6.0 %), 319 (6.4 %), 272 (5.4 %), 298 (6.0 %), 518 (10.3 %), 408 (8.1 %), 804 (16.1 %), and 1553 participants (31.0 %) with IMD levels from 1 to 10, and 362 participants (7.2 %) missing IMD information.

3.2. Model establishment

In the context of GP assessing and referring patients with RMD, patients' Presenting Condition Description (PCD) is normally generated during consultation with GP, Medical History (MH) is available from EPR, while Blood Test Data (BTD) will be ordered if necessary and is less frequently available during patients' first visit to GP for RMD problems. BTD usually becomes available after patients visit their GP or are referred to a specialist. As shown in Fig. 1, the proposed machine learning framework consists of an attention-based multimodal feature representation fusion network, a BTD-based classification module, and a late fusion module to fuse predictions from preceding submodules. To enhance the adaptability of the proposed models to patients with diverse data modalities, a hierarchical fusion process is implemented and further detailed in **Fig. S2** of **Supplementary Material**. This method can ensure robust and flexible predictions tailored to the availability of input data. In practice, for patients with only PCD and MH data, the attention-based fusion model directly provides risk stratification predictions. However, for patients with additional BTD data, the GBM-based model is utilized, and its predictions are then fused with those of the attention-based model using a Bayesian approach.

The attention-based feature fusion network has been proposed as the multimodal fusion method to model the underlying relationships between unstructured presenting condition description data and semi-structured medical history data. Specifically, the multimodal fusion network consists of two transformers-based language model encoder networks that serve as the feature extractors to extract feature representations from presenting condition description data and medical history data, and an attention-based feature fusion network that effectively integrates feature information and their underlying relationships extracted from the two data types. A detailed description of the attention-based multimodal fusion is included in Section 3.2.2.

The BTD data based classification module is proposed to differentiate IC and NIC using patient's blood test results data. In this study, the gradient boosting machine (GBM) model is utilized due to its capability in handling data with missing values, and thus no further data imputation steps would be required. The LightGBM is chosen to implement this method because it provides a leaf-wise algorithm that can reduce loss

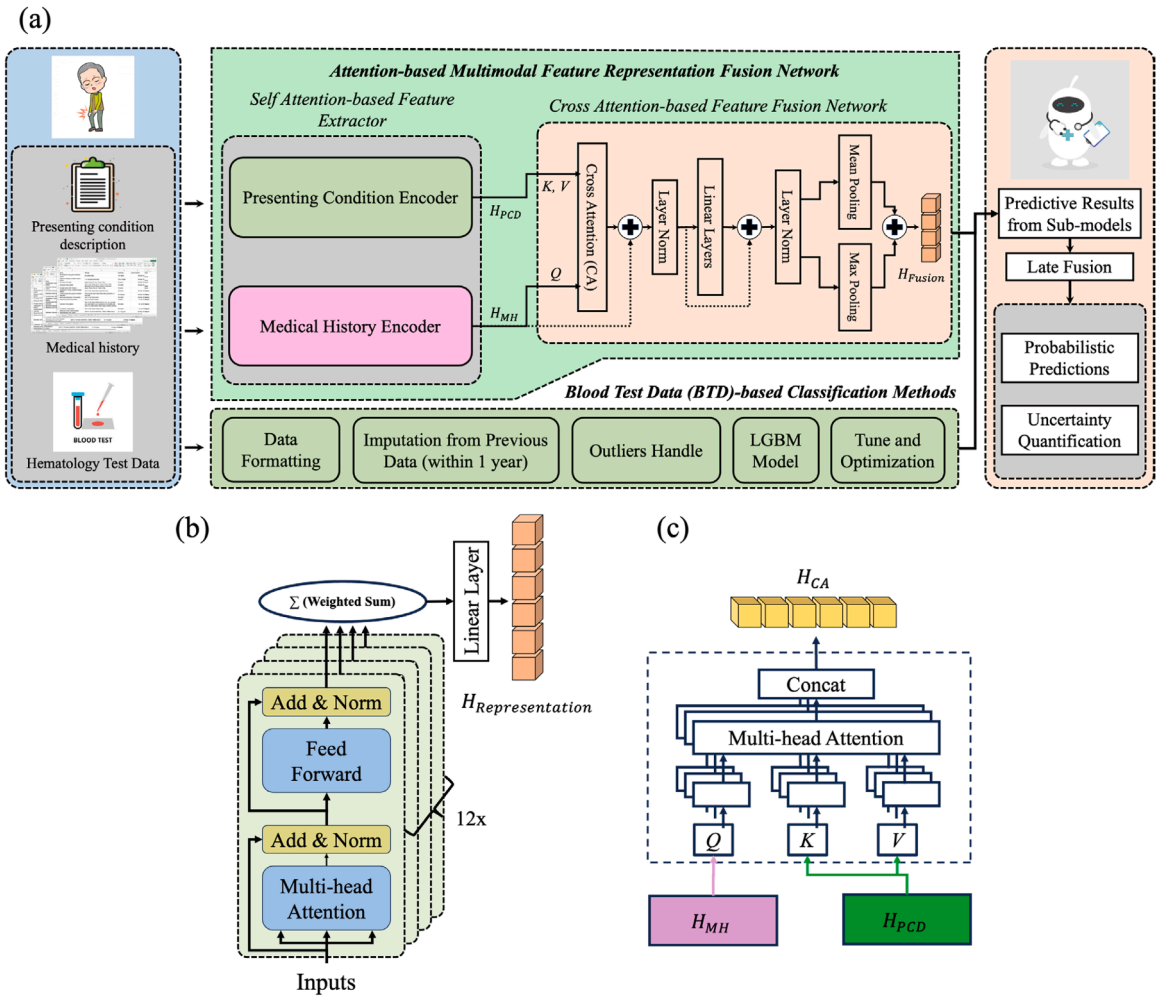

Fig. 1. Framework of the proposed methods: (a) Overall architecture of the system; (b) Transformers-based multi-layers fusion network is the fundamental structure of the Presenting Condition Encoder and Medical History Encoder in (a); (c) Attention-based feature fusion network is the initial network structure of the cross attention-based feature fusion network in (a).

and improve accuracy and support faster training speed as well as higher efficiency.

Most patients have PCD and MH data simultaneously during referrals. Only a small portion of patients have BTD, PCD and MH all simultaneously. There is insufficient blood test data to train the attention-based fusion network with the corresponding presenting condition descriptions and medical history data simultaneously. Therefore, the ensemble learning-based late fusion module is proposed to fuse predictions from the attention-based fusion network submodule and the BTD data based classification submodule. Ensemble learning is a commonly used method to fuse predictions to improve the model's accuracy and boost the generalizability of the model. In this study, a Naïve Bayes model was utilized to ensemble the predictions from the attention-based feature fusion network (based on language models using presenting conditions and medical history) and the GBM-based classification method (based on blood results). A comprehensive elaboration on the ensemble learning based fusion module can be found in the Section 3.2.3.

3.2.1. Transformer-based language model networks

The transformer-based language model networks consist of two sub encoder networks, including a PCD encoder to extract feature representation for patients having PCD data from GP referral letters and a medical history encoder for patients having clinical information summary (Fig. 1 **(b))**. These two encoder networks have the same transformer-based structures that are based on Bidirectional Encoder Representations from Transformers (BERT) [59]. It has been proven that different layers of the BERT model can extract features of the inputs from various levels [60]. Therefore, we dynamically fused outputs of multiple intermediate layers of BERT with learnable weights. Specifically, after the transformation, the dense tensor representations H_{PCD} and H_{MH} would be extracted by PCD encoder and MH encoder, which will be fed into the attention-based fusion network to further extract the multi-modal features between different data modalities.

3.2.2. Attention-based feature representation fusion network

The proposed Attention-Based Feature Representation Fusion Network (AFRFN) integrates diverse data modalities, such as PCD and MH modalities, into a unified representation to enhance predictive performance. By leveraging an attention mechanism, the network dynamically extracts features from different modalities, ensuring that the most informative features could be captured, such as complementary and supplementary features across various modalities. Specifically, through independently encoding separate modality using language model-based encoders, the textual input data is transformed into high-dimensional feature vectors. These feature vectors are then fed into the subsequent attention-based fusion layers, which produce the final representation vector by fusing features extracted from the different language model encoders.

As shown in Fig. 1, the attention layers calculate the attention values of the PCD representation H_{PCD} over the MH representation H_{MH}, based on the Eqs. (1) and (2). After the attention transformation, we will get the attention tensors $Attn_{PCD,\ MH}$ based on PCD and MH representation outputted from the PCD encoder and MH encoder as described in the above section. Notably, a multi-heads attention mechanism will be applied into the attention calculation.

$$\begin{aligned} Head_M(Q,K,V) &= Concat(a_1,a_2,\ldots,a_h)W^O \\ a_i &= Attention\left(QW_i^Q, KW_i^K, VW_i^V\right) \end{aligned} \tag{1}$$

where $Head_M(Q,K,V)$ is the multi-heads attention tensor, a_i is the attention tensor of *i-th* head, which is calculated by Eq. (2).

The attention mechanism is detailed in study [61], and the computation process is described in Eq. (2), which calculates the weighted summation of values V on the basis of similarity between keys K and queries Q. Distinctively, in the process, H_{MH} is employed as the query, and H_{PCD} as the key and value to model the triplet $(H_{MH}, H_{PCD}, H_{PCD})$, while in self-attention, the same tensor is employed as the query, key, and value, to model the tensor triplet.

$$Attention(Q,K,V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2}$$

where Q, K, and V are query tensor, key tensor, and value tensor, respectively, and $d_k = d_{model}/h$, where h is the heads of the attention layers.

Following the multi-modal attention feature fusion network, a multiple layer feed forward network is applied to the outputs of the attention layers $Attn_{PCD,\ MH}$ to further extract features.

$$H_{Fusion} = MFF\left(Attn_{PCD,\ MH}\right) \tag{3}$$

where $Attn_{PCD,\ MH}$ is attention tensor of the final hidden layer of attention-based feature fusion network, $MFF(*)$ represents the multi-layer feed-forward network and H_{Fusion} is the final output of the network. It is worth noting that different layers have been introduced into the network, including Norm Layer, Residual Connection, Max and Mean Pooling Layers.

At the end of the multimodal attention-based model, a linear layer is applied to get the final classification. The calculation process is described in Eq. (4).

$$output = Linear(H_{Fusion}) \tag{4}$$

where H_{Fusion} is the output tensor of multimodal fusion network, and $Linear(*)$ represents the final linear classification layer.

3.2.3. Ensemble learning method

Ensemble learning is the commonly used methods to improve the model's accuracy and robustness, and shows great potential in boosting the generalization capabilities of the model. Traditional ensemble learning methods, such as bagging, boosting, and stacking [62,63], aim at combining the predictions of different single classifiers to get a better performance compared with the single classifiers. In our study, a Naïve Bayes-based ensemble learning method has been developed to fuse probabilistic predictions of preceding models for patients who have PCD, MH, and BTD simultaneously, as shown in Fig. 1.

3.3. Detecting unreliable RMD risk stratifications using conformal prediction

To ensure the predictions are robust and GPs or clinicians can understand the uncertainty associated with the point prediction of RMD risk stratification, especially if any individual patient attributes are out of training dataset distribution. We used a conformal prediction [64] based method to generate a prediction set of RMD risk stratification (e.g. either IC or NIC) with a predefined confidence level [35]. If the prediction does not reach the predefined confidence level, an empty prediction can be made, or, if the prediction set associated with an RMD risk point prediction (e.g., IC or NIC) is too large for the prediction to be informative, a multiple prediction can be made. The RMD risk point prediction with corresponding multiple or empty conformal prediction sets can be flagged for human intervention. The conformal predictor is used with the RMD risk stratification model to detect unreliable RMD risk predictions and guarantee the error rate is bounded by a pre-specified level [65]. A detailed description of the conformal prediction-based method to detect unreliable RMD risk prediction is described below.

For any given confidence level, conformal predictor can adjust any classifier's point predictions to predictive sets. The computing process of the conformal predictor includes a calibration step (Algorithm 1) and an inference step (Algorithm 2**)**. The first step is to calculate the $\hat{q}$ value as described in Algorithm 1 using the calibration set, and the second step is

Algorithm 1
Conformal prediction calibration step.

Inputs: Pre-defined error rate α, and calibration set (X^C, Y^C)
procedure *Conformal Calibration*(α, X^C, Y^C)
  $S = \left\{ s_i = 1 - \widehat{f}(X_i^C)_{Y_i^C}, i \in \{1, \ldots, n\} \right\}$ § $\widehat{f}$ is any classifier
  $\widehat{q} \leftarrow Q(S, q)$, $q \leftarrow \frac{\lceil (n+1) * (1-\alpha) \rceil}{n}$ § The quantile function Q
  return $\widehat{q}$
Outputs: $\widehat{q}$

Algorithm 2
Conformal prediction inference step.

Inputs: $\widehat{q}$ calculated from calibration step, and test set (X^T, Y^T)
procedure *Conformal Predictor*$(\widehat{q}, X^T, Y^T)$
  $C \leftarrow \{\ \}$
  for $i \in \{1, \ldots, m\}$ **do**
    $C(X_i^T) = \left\{ y : \widehat{f}(X_i^T)_y \geq 1 - \widehat{q}, y \in \{1, \ldots, K\} \right\}$ § classifier $\widehat{f}$
    $C \leftarrow C(X_i^T)$
  return C
Outputs: The prediction set C with $1 - \alpha$ coverage

to infer the prediction set for each instance in the test set as shown in Algorithm 2.

Conformal prediction is a mathematical framework that can be used together with any machine learning system or prediction model to guarantee the error rate is bounded by a pre-specified level [65], and the conformal coverage could be guaranteed by Theorem 1.

Theorem 1. Let (X^C, Y^C) and (X^T, Y^T) be independent and identically distributed (*i.i.d.*), and define $\widehat{q}$ as in step 3 of Algorithm 1 and (X^T, Y^T) as in Algorithm 2. Then, the following formula holds:

$$P\left(Y^T \in C(X^T)\right) \geq 1 - \alpha$$

Where α is the pre-specified error rate, $C(X)$ is an uncertainty set function, which maps a feature vector $X \in R^d$ to a subset of $\{1,2,\ldots,K\}$, K is the total number of labels, and a detailed statement of this theorem could be found in [56].

Unlike the single predictive results (point prediction, e.g., IC or NIC) made by the machine learning model, conformal predictor will output a prediction set with a pre-defined confidence level rather than a single prediction point by the model, which can be used to measure the uncertainty of the model's prediction. For classification problems, this prediction region corresponds to a set of labels, a multilabel prediction. Based on [35], four kinds of prediction sets are defined in this study, including multiple, single, error and empty predictions, which are defined in **Table S1** in **Supplementary Material**. To explain the definitions of various prediction sets, we clarify these concepts based on a real label of IC, which makes the definitions clearer. If the prediction does not reach this confidence level, an empty prediction can be made. If the prediction region associated with a point prediction is too large for the prediction to be informative, a multiple prediction can be made. The corresponding empty and multiple prediction can be flagged for human intervention. The conformal predictor can thus function as a quality control system for ensuring that only reliable predictions are made.

3.4. Quantifying multimodal contributions

Due to the inherent black-box nature of most deep learning-based models, its lack of transparency and interpretability poses challenges in providing clear explanations to physicians, impeding the practical integration of these models in real-world clinical decision support systems [66]. To address these concerns, eXplainable AI has been extensively explored within the context of unimodal machine learning methods. However, multimodal eXplainable AI (MXAI) remains relatively underexplored due to several challenges [67]: (1) quantifying the contributions of individual modalities to identify the most significant modality on the model's overall predictions; (2) identifying the most salient features within each modality from the multimodal level to uncover the potential complementary or supplementary relationships among various modalities; (3) detecting causal relationships in the model's input-output data to facilitate the generation of explanations that are more intuitive and comprehensible for human users.

Apparently, a key advantage of MXAI, compare with unimodal XAI, is to support explanations using multiple modalities that are complementary and cover more aspects of explainability [67]. Furthermore, XMAI can help clinicians trace back disease predictions. This traceability not only builds trust but also supports transparent early detection recommendations [68]. Enhancing the interpretability of these models is particularly important in high-stakes applications like healthcare. Therefore, MXAI stands out as a challenging yet promising research area that attracted researchers' extensive attention.

To address above challenges, this study employs multimodal Shapley values (MM-SHAP) [69] to explain the predictions made by the proposed multi-stage multimodal fusion network. MM-SHAP, formally defined in Eq. (5), provides a systematic approach to quantify the contribution of each data modality to the model's final predictions, which are critical for improving transparency of black-box models, particularly in high-stakes applications where understanding the decision-making process is essential for real-world deployment.

$$MM - SHAP_i = \frac{\sum_{j=1}^{n_i} \left| \Phi_i^j \right|}{\sum_{i=1}^{I} \sum_{j=1}^{n_i} \left| \Phi_i^j \right|} \tag{5}$$

Here, Φ_i^j represents the Shapley value associated with the input entry j within the given input modality i, which follows formula Eq. (6), and n_i is the input length of modality i.

$$\Phi_i^j = \sum_{S \subseteq M \setminus \left\{ M_i^j \right\}} \frac{val\left(S \cup \left\{ M_i^j \right\}\right) - val(S)}{\gamma} \tag{6}$$

Where, $\gamma = \frac{|M|!}{|S|!(|M|-|S|-1)!}$ is the normalizing factor that accounts for all possible combinations of choosing subset S. The model's input sequence, denoted as $M = \{M_I^J\}$, consists of $|M| = \sum_{i=1}^{I} n_i$ input entries, where i and j index the input modality, and input entry corresponding to the specific data modality.

In this study, we used MM-SHAP to interpret model's predictions made by the proposed multi-stage multimodal fusion network. This approach provides insights into the model's decision-making process, and presents clinicians with interpretable prediction results by highlighting key words and variables. Furthermore, this method can identify primary data modalities that contribute most to the predictions.

3.5. Model training and performance evaluation

A comprehensive description about how to train the proposed models is elaborated in **Section S3** of the **Supplementary Material**. Specifically, the attention-based multimodal feature representation fusion network designed for patients with PCD and MH data is trained following the tactics in the **Section S3.1**, and the GBM-based classifier proposed for patients with only BTD is trained as described in **Section S3.2**. The Bayesian model, developed for patients with complete data, is trained using the strategies detailed in **Section S3.3**. Moreover, the training procedures for the proposed conformal predictors are elaborated in **Section S3.4**.

Additionally, the performance of the proposed models was evaluated by various evaluation measures, including the sensitivity, specificity, accuracy, AUC, ROC curve, and G-Mean. Furthermore, two metrics were

proposed to measure the performance of the conformal predictor, including the empirical coverage and error flag rate, which are formulated as Eq. (7). Simply, $C_{Empirical}$ represents the percentage of single and multiple predictions generated by conformal predictor over whole testing set given a pre-defined confidence level (or error rate); R_{Flag} denotes the proportion of multiple and empty prediction sets flagged by the conformal predictor over model's incorrect point predictions. Generally, an optimal conformal predictor features higher empirical coverage, and can be used to identify unreliable predictions where machine learning model made wrong predictions. However, the multiple predictions would increase given a higher confidence level, subsequently leading to an increase in uncertain predictions. Therefore, in practical applications, confidence level should be optimized to balance the error flag rate and empirical coverage, and then clinician's intervention could be introduced to prioritize the review of these patients.

$$C_{Empirical} = \frac{N_{p_{cp}=\{Single\}\cup\{Multiple\}}}{N}$$
$$R_{Flag} = \frac{N^{p_{cp}=\{Multiple\}\cup\{Empty\}}_{l_{Real}\neq p_{Model}}}{N_{l_{Real}\neq p_{Model}}} \quad (7)$$

where N is the size of test set, l_{Real}, p_{Model}, and p_{cp} are patient's real diagnosis, machine learning model's predictions, and the prediction sets of the conformal predictor, $N_{p_{cp}=\{Single\}\cup\{Multiple\}}$ is the total patients that the conformal predictor assigns single and multiple prediction set, $N^{p_{cp}=\{Multiple\}\cup\{Empty\}}_{l_{Real}\neq p_{Model}}$ is the total patients that machine learning model makes wrong prediction but the conformal predictor flags as multiple prediction set, and $N_{l_{Real}\neq p_{Model}}$ is the total patients where machine learning model makes wrong predictions.

4. Experimental results

4.1. Data preprocessing

The inputs of the proposed model (Fig. 1) involve three data modalities, including the unstructured PCD input, semi-structured MH input, and structured BTD input. Specifically, the inputs of the transformer-based language model networks are unstructured PCD and semi-structured MH data. For the unstructured PCD data, the preprocessed texts were truncated (longer than 512 tokens) or padded (shorter than 512 tokens) to a fixed length of 512 tokens before feeding to the language models because the majority of data's token lengths are below this threshold, which can ensure that critical clinical information is preserved for most patients. Slightly different from the PCD data, the MH data involves two steps before feeding to the language models. The first step is to rank different records by using timestamps, and then concatenate main contents in different sections into one text (e.g., drug in Medication History; problem in Problems History; description in Allergies History and Social Context History), and the second step is the same as the PCD data as described precedingly. For the BTD input, the preprocessed data was directly fed to the GBM model with 29 features. The detailed data preprocessing methods are provided in **Section S2** of **Supplementary Material**.

4.2. Comparison of different baseline models

The performance of the proposed Attention-based Feature Representation Fusion Network based on various pre-trained BERT language models has been compared against a range of baseline models, such as RoBERTa, Llama 3.1–8B, Llama 3.1–70B, Qwen 2.5–7B, Qwen 2.5–72B. As shown in Table 2, although the baseline large language models (Llama-3.1–8B, Llama-3.1–70B, Qwen2.5–72B) achieved the sensitivity of 0.97, they consistently demonstrate very low specificity and low G-Mean values meaning not acceptable in practice. Notably, our method based on the Bio-ClinicalBERT model achieved the best G-Mean of 0.75 compared with other language models. This suggests that using Bio-ClinicalBERT and Attention-based Feature Representation Fusion Network can achieve better performance, and our method is more suitable for practical clinical applications.

Table 2
Comparison of the proposed attention-based fusion model with various baseline language models for patients having PCD and MH data.

Models	Specificity	Sensitivity	Accuracy	AUC	G-Mean
RoBERTa	0.78	0.56	0.71	0.76	0.66
Llama-3.1–8B	0.22	0.97	0.46	-	0.47
Llama-3.1–70B	0.16	0.97	0.42	-	0.39
Qwen2.5–7B	0.24	0.94	0.47	-	0.48
Qwen2.5–72B	0.24	0.97	0.47	-	0.48
AFRFN-BERT-Base	0.79	0.61	0.73	0.74	0.69
AFRFN-BioBERT	0.73	0.66	0.70	0.77	0.69
AFRFN-ClinicalBERT	0.64	0.50	0.60	0.61	0.57
AFRFN-Bio-ClinicalBERT	0.68	0.83	0.73	0.79	0.75

Note: AFRFN (Attention-based Feature Representation Fusion Network) is trained using different pre-trained language models, including BERT-Base (AFRFN-BERT-Base), BioBERT (AFRFN-BioBERT), ClinicalBERT (AFRFN-ClinicalBERT), Bio-ClinicalBERT (AFRFN-Bio-ClinicalBERT).

Additionally, Table 3 demonstrates a detailed comparison of different machine learning models for patients with BTD data. In the model training process, missing values were imputed using mean values, and all features were normalized using a standard scaler. Among the evaluated models, the gradient boosting machine exhibited superior overall performance, achieving the highest metrics across most evaluation categories. Notably, the GBM without data imputation and scaling achieved a slightly better G-Mean compared to GBM with data imputation and scaling, which suggests the robustness of GBM models compared to other machine learning approaches.

Furthermore, Table 4 provides a comprehensive comparison of various fusion methods for patients with complete PCD, MH, and BTD data. Notably, the Support Vector Machine model achieves the highest AUC of 0.92 and specificity of 0.98 in identifying NIC patients, its sensitivity remains comparatively low for detecting IC patients. The Gaussian Naïve Bayesian model demonstrates the best overall performance with the highest G-Mean 0.89, AUC 0.92, accuracy 0.90, specificity 0.91, and sensitivity 0.88, making it the most balanced method in terms of differentiating NIC and IC patients.

4.3. Comparison based on data availability during referrals

Furthermore, we compared the overall performance between language model and attention-based fusion network modeling on patients

Table 3
Comparison of different machine learning methods for patients with BTD data.

Models	Specificity	Sensitivity	Accuracy	AUC	G-Mean
Logistic Regression	0.70	0.53	0.66	0.63	0.61
Support Vector Machine	0.60	0.65	0.61	0.62	0.62
Random Forest	0.71	0.55	0.67	0.73	0.62
Gaussian Naïve Bayesian	0.69	0.56	0.66	0.67	0.62
Decision Tree	0.66	0.58	0.64	0.71	0.62
Linear Discriminant Analysis	0.72	0.45	0.65	0.62	0.57
Quadratic Discriminant Analysis	0.62	0.58	0.61	0.63	0.60
Gradient Booting Machine (with data imputation)	0.75	0.69	0.71	0.78	0.72
Gradient Booting Machine (without data imputation)	0.69	0.76	0.71	0.78	0.73

Table 4
Comparison of different fusion methods for patients with complete data.

Models	Specificity	Sensitivity	Accuracy	AUC	G-Mean
Decision Tree	0.81	0.53	0.74	0.67	0.66
Gradient Booting Machine	0.83	0.53	0.76	0.85	0.66
Support Vector Machine	0.98	0.47	0.86	0.92	0.68
Random Forest	0.94	0.41	0.81	0.85	0.62
Linear Discriminant Analysis	0.91	0.65	0.84	0.92	0.77
Gaussian Naïve Bayesian	0.91	0.88	0.90	0.92	0.89

having both PCD and MH data, and Naïve Bayes-based ensemble modeling on patients having PCD, MH, and BTD data simultaneously (ROC in Fig. 2). As shown in Table 5, the proposed attention-based fusion network for patients having PCD and MH data can achieve the specificity, sensitivity, accuracy, AUC, and G-Mean values of 0.68, 0.83, 0.73, 0.79, and 0.75, in identifying NIC and IC patients. Furthermore, our experimental results indicate improved performance by further including the BTD data in modeling an ensemble method, achieving the specificity, sensitivity, accuracy, AUC, and G-Mean values of 0.91, 0.88, 0.90, 0.92, and 0.89.

4.4. Reducing RMD risk stratification errors by detecting unreliable predictions

The conformal predictor has been used to identify uncertain predictions of the proposed model. Table 6 (**Table S4** in **Supplementary Material**) showed the comparison results between prediction sets of the conformal predictor with 95 % confidence and point predictions of the machine learning based RMD risk stratification model. The conformal predictor has made 23.38 % predictions as single predictions (IC or NIC) correctly, 7.36 % predictions as error single predictions, and 69.26 % predictions as multiple predictions, indicating that the predictions are uncertain, and the model cannot distinguish between several possible class labels at the pre-defined confidence. Out of the 70 false predictions (30.3 % of all predictions) made by the standalone machine learning risk stratification model, 53 of them (75.71 %) (**Table S5** in **Supplemental Material**) have been flagged by the conformal predictor as multiple predictions, meaning those predictions are of high uncertainty and needs clinicians to have a second review.

Table 5
Performance of RMD risk stratification model with different types of data.

Models	Data types	Specificity	Sensitivity	Accuracy	AUC	G-Mean
AFRFN Model	PCD, MH	0.68	0.83	0.73	0.79	0.75
GBM-based Model	BTD	0.69	0.76	0.71	0.78	0.73
Ensemble Model	PCD, MH, BTD	0.91	0.88	0.90	0.92	0.89

Note: This table reports the performance of the proposed models using different data types during referral. Abbreviations are: Presenting Condition Description (PCD), Medical History (MH), Blood Test Data (BTD). Attention-based Feature Representation Fusion Network (AFRFN) is developed for patients with PCD and MH data, GBM-based model is developed for patients with BTD data, and the ensemble model is for patients with PCD, MH, and BTD data.

4.5. Enhancing model interpretability by quantifying multimodal contributions

Fig. 3 illustrates the SHAP explanations made by the proposed multistage multimodal machine learning model for a real patient with the inflammatory condition. The model accurately predicts the patient's early inflammatory condition, and MM-SHAP analysis reveals that PCD, MH, and BTD inputs separately contribute 25.7 %, 51 %, and 23.3 % to the patient-level prediction. This indicates that MH data plays a key role in influencing the model's prediction compared to the PCD and BTD modalities for this patient.

Additionally, the SHAP analysis identifies key contributing features across each data modality. In the PCD data, terms such as *painful swollen foot left, colchicine*, and *a trial of prednisolone* are highlighted as positive contributors. Similarly, in the MH data, keywords including *Co-codamol, Flucloxacillin, Prednisolone, Colchicine, Omeprazole*, and *Simvastatin* are identified as positive contributors, whereas *Ex-smoker* is highlighted as a negative contributor. For the BTD data, blood indicators such as *Eosinophil count, Haemoglobin*, and *Platelet count* are identified as positive contributors, while *rheumatoid factor* is noted as a negative contributor.

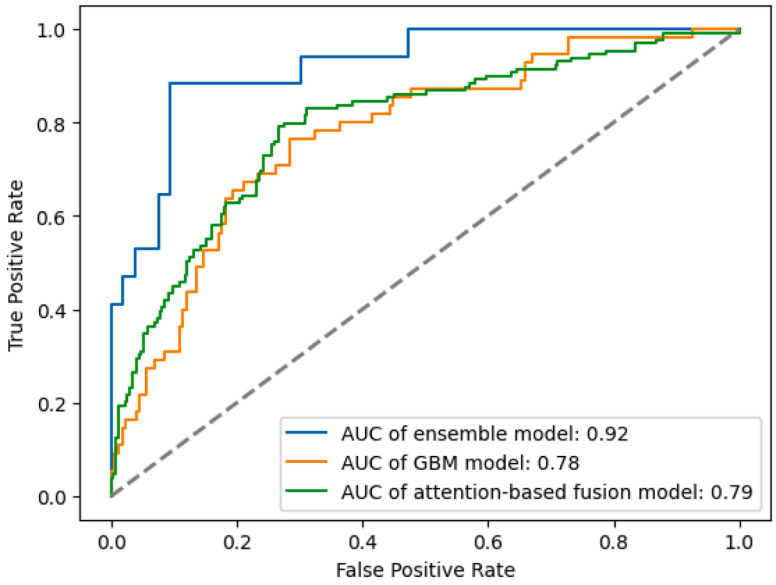

Fig. 2. ROC curve of the proposed models, which visualize the ROC curves of the proposed models using different data types, including patients separately having PCD and MH data (language model and attention-based fusion model), BTD data (GBM model), as well as PCD, MH, and BTD data (ensemble model).

Table 6
Result of prediction regions of the conformal predictor and point predictions of the machine learning risk stratification model for patients having PCD and MH data.

Subgroups	NIC	IC	All
Conformal prediction regions (Confidence 95 %)			
Error, n (%)	15 (8.62 %)	2 (3.51 %)	17 (7.36 %)
Empty, n (%)	0 (0)	0 (0)	0 (0)
Single predictions, n (%)	30 (17.24 %)	24 (42.11 %)	54 (23.38 %)
Multiple predictions, n (%)	129 (74.14 %)	31 (54.39 %)	160 (69.26 %)
Machine learning point predictions			
False predictions, n (%)	62 (35.63 %)	8 (14.04 %)	70 (30.3 %)
True predictions, n (%)	112 (64.37 %)	49 (85.96 %)	161 (69.7 %)
Machine learning point prediction + conformal prediction regions			
False predictions flagged by multiple predictions, n (%)	47 (75.81 %)	6 (75 %)	53 (75.71 %)

Note: The early risk stratification performance of the RMDs is presented in both prediction regions by conformal predictor and point predictions by machine learning model. The results are reported at a confidence level of 95 %. Labels (NIC and IC) are included in the prediction region if their confidence is higher than a pre-defined confidence (95 %). The error prediction represents the portion of true labels not included in the prediction region. A multiple prediction indicates that the prediction is uncertain, and the model cannot distinguish between several possible class labels at the pre-defined confidence. An empty prediction is where the model could not assign any label, typically meaning that the example is very different from the data the model was trained on. False predictions flagged by multiple predictions are the percentage of false predictions made by the machine learning model and simultaneously with multiple prediction regions made by the conformal predictor.

These findings provide potentially valuable insights into the model's underlying decision-making processes. By uncovering how the diverse data sources contribute to its predictions, the analysis can help improve the model's interpretability and guide clinicians toward more informed and accurate diagnoses. This also sheds light on the further improvement of explainability methods to ensure its predictions align more closely with clinical expertise and reasoning.

5. Risk stratification decision support in real-world RMD referral workflow

Fig. 4 illustrates how our model supports RMD referral processes in the real-world through risk stratification decision support. When a patient sees and consults a GP, the GP will type the presenting condition of the patient for the current visit into the system. The decision support system will automatically process the presenting condition, retrieve medical history and available blood test results data from Electronic Patient Record. These data are subsequently fed into the proposed models to generate three key decision support information: 1) RMD risk stratification: risk of having IC and NIC; 2) key risk factors contributed to the RMD risk prediction; 3) level of uncertainty of the predicted RMD risk. Those information will be presented in a way that end users such as GPs and secondary care clinicians can understand easily. Take the patient mentioned in Section 4.5 as an example, based on this patient's PCD, MH and BTD data, our model predicts the patient with a 51 % probability of having IC and have analyzed the key contributing factors and uncertain level. As shown in Fig. 4, the decision support will also provide a breakdown of contributions of different data modalities to the prediction, attributing 25.7 %, 51 %, and 23.3 % of PCD, MH, and BTD modalities respectively in this example. Furthermore, the decision support will highlight key words in the PCD and MH data that significantly contribute to the risk prediction, such as the *painful swollen foot left, colchicine*, and *a trial of prednisolone* from PCD data, and *Co-codamol, Flucloxacillin, Prednisolone, Colchicine, Omeprazole*, and *Simvastatin* from MH data. Simultaneously, it ranks blood test results from high to low by their relative importance to the predicted risk, such as the *Eosinophil count, Haemoglobin*, and *Platelet count*. Additionally, the decision support will output the level of uncertainty of the prediction given the confidence level. In this example, the uncertainty of patient of having IC is low.

Based on all the decision support information of RMD risk stratification, multimodal explanations, and uncertainty analysis, GP will make the decision to refer the patient. If the patient is likely to have IC based on all the information and the decision support, then the GP will refer the patient to rheumatology specialists at the secondary hospital. If the patient is likely to have NIC based on all the information, then the GP will refer the patient to physiotherapy or orthopaedics specialists.

If the uncertainty level is high for the RMD risk stratification, the patient will be flagged for GP for a second review or suggest seeking Advice and Guidance [36] from the specialist. This can further enhance the safety of the decision support by mitigating false predictions by the model. For example, if the machine learning model incorrectly suggests a patient with non-inflammatory condition. However, conformal prediction identifies the patient belonging to both non-inflammatory and inflammatory condition categories, which means the RMD risk stratification prediction is uncertain and unreliable and thus needs a second review with more cautious assessment.

6. Conclusion and discussion

6.1. Strength

In this study, we developed machine learning models to risk stratify RMD diseases by identifying and differentiating inflammatory conditions and non-inflammatory conditions using data available during referrals. Our method can accommodate patients with different data types during referrals including unstructured presenting condition description, semi-structured medical history, and structured blood test results. With the presenting conditions and medical history information which are normally available when a patient visits a GP, the model alone can identify and differentiate IC and NIC with 0.68 specificity, 0.83 sensitivity, and 0.73 accuracy, which is significantly higher than GP [70]. If the patients have blood test results available together with PCD and MH, our model can achieve 0.91 specificity, 0.88 sensitivity, and 0.90 accuracy. Our models are developed and validated using data from 128 GP practices. We also experimented using different existing pre-trained language models as feature extractors in developing the risk stratification model. Experimental results illustrated that the proposed attention-based feature fusion model achieved better performance when using Bio-ClinicalBERT base model, as shown in Table 2.

Unreliable predictions can occur when a machine learning model is presented with data it has not been exposed to during training. We demonstrate the use of conformal prediction to detect unreliable predictions to ensure robust prediction and patient safety when implementing the machine learning model in real-world referrals. For machine learning predictions with patients' presenting conditions and medical history, 75.71 % of false predictions can be flagged by the conformal predictor as unreliable predictions, which can potentially further reduce the total false prediction rate to 7.36 % (**Table S5** in **Supplementary Material**). For machine learning predictions with patients' presenting conditions, medical history, and blood test results, 66.67 % false predictions have been flagged, potentially reducing the total false prediction rate to 5.71 % (Table 7 and **Table S7** in **Supplementary Material**). The predictions being flagged as unreliable ones can be further assessed by human clinicians. This means combining the point prediction from the machine learning risk stratification model, with conformal prediction sets, and human assessment of unreliable cases can improve the accuracy of identifying IC and NIC of RMD patients. For example, assuming unreliable predictions being flagged can be corrected by humans, a combination of our models (machine learning

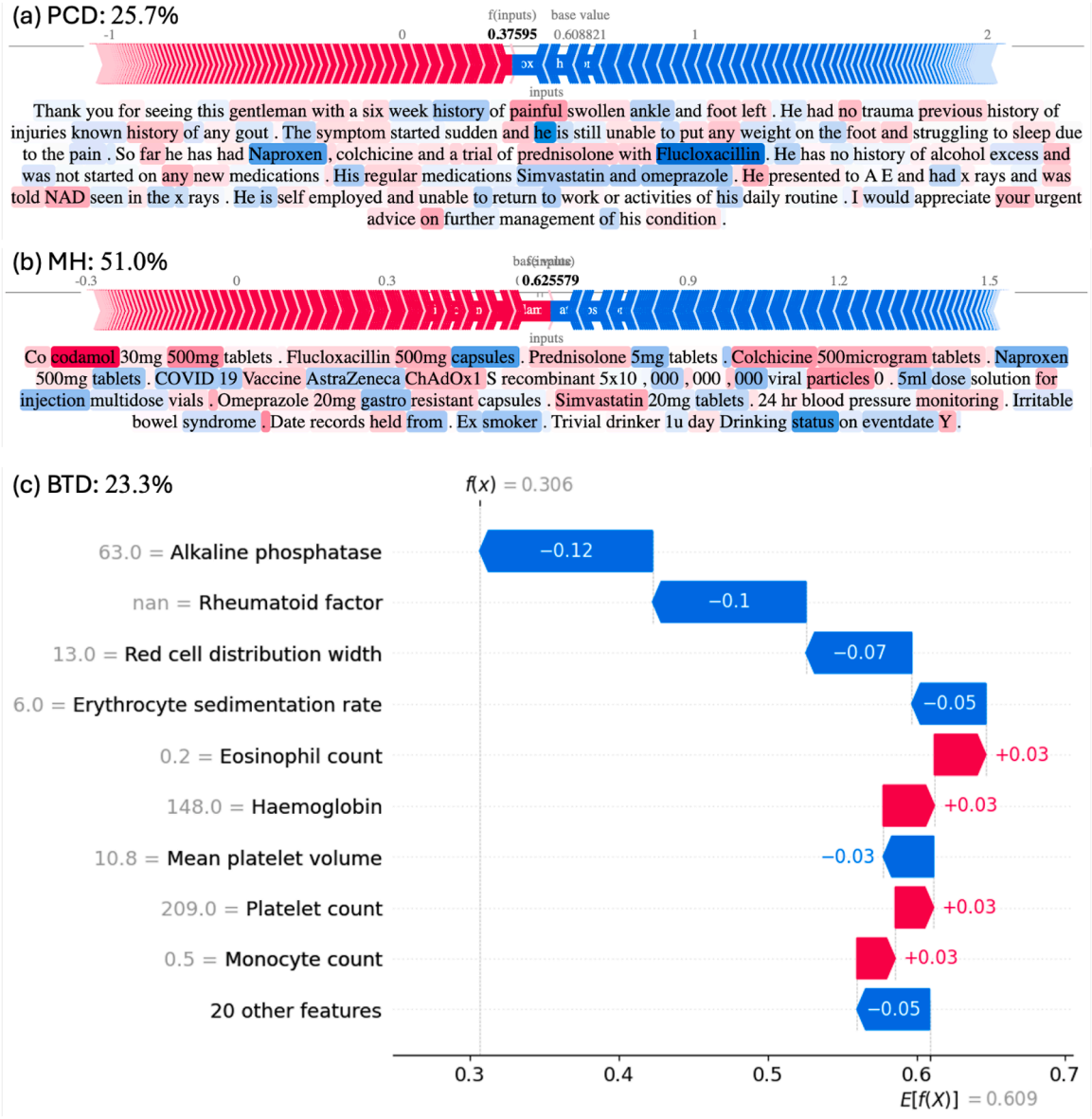

Fig. 3. Measurement of multimodal contributions by using MM-SHAP. (a) SHAP explanation of PCD input; (b) SHAP explanation of MH input; (c) SHAP explanation of BTD input.

and conformal predictor) and human assessment can achieve 0.91 specificity and sensitivity 0.96 using only presenting condition description and medical history.

Our models can be deployed in the hospital and used as decision support systems for doctors (clinicians and GP) to detect and differentiate patients with IC or NIC so that patients can receive the right treatment faster. With the capability to analyze symptoms and medical history in free text, our model can also be used by patients directly describing the conditions themselves so that they can be advised to visit the right specialist well in advance to receive the right treatment.

6.2. Limitations

Our existing method involves manually processing raw referral data from PDF/Word by extracting presenting condition and medical history information from raw referral letters into a machine-readable format so that the relevant information can be fed into our model for training and testing. However, in the real-world application, real-time decision support enabled by our model is needed for GPs and clinicians during referrals. This means an automatic data preprocessing method to extract relevant presenting condition and medical history information from free-form referral letters to model inputs accurately is needed, which will be developed in our future work for real-world implementation.

Although referral data from 128 GP practices were used in model development and validation providing unique external validation opportunities, there are still limitations to model generalizability associated with inherent bias in the healthcare system. The data is collected from patients who visited GPs and we may miss data for patients who have similar problems but do not have access to GPs. In the future, more data from more patients and wider regions will be collected to further improve the generalizability of the model. To this end, we will be collaborating with more GPs from various regions in the UK so that our method can be further optimized and validated with large scale datasets from broader populations. Furthermore, we will also conduct prospective trials to evaluate the effectiveness and safety of our method in the real world.

Additionally, further work will assess the fairness of our model among patient subgroups and compare potential selection bias between

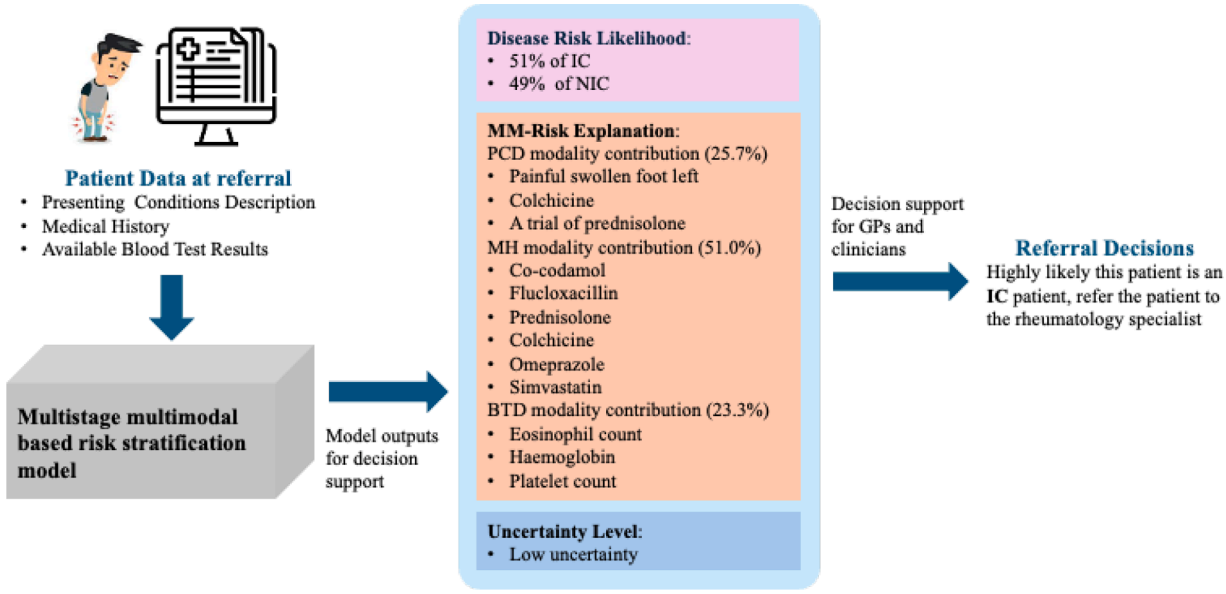

Fig. 4. RMD Risk stratification workflow in referral processes after incorporating the proposed models.

Table 7
Result of prediction regions of the conformal predictor and point predictions of the machine learning risk stratification model for patients having PCD, MH, and BTD data.

Subgroups	NIC	IA	All
Conformal prediction regions (Confidence 95 %)			
Error, n (%)	0 (0)	2 (25.00 %)	2 (5.71 %)
Empty, n (%)	0 (0)	0 (0)	0 (0)
Single predictions, n (%)	19 (70.37 %)	2 (25.00 %)	21 (60 %)
Multiple predictions, n (%)	8 (29.63 %)	4 (50.00 %)	12 (34.29 %)
Machine learning point predictions			
Error, n (%)	4 (14.81 %)	2 (25 %)	6 (17.14 %)
Correct, n (%)	23 (85.19 %)	6 (75 %)	29 (82.86 %)
Machine learning point prediction + conformal prediction regions			
False predictions flagged by multiple predictions, n (%)	4 (100 %)	0 (0)	4 (66.67 %)

Note: Same as the note for Table 6.

human and our model. Bias correction methods will be developed in future work to mitigate any potential bias of the model for real-world application [71,72].

6.3. Conclusions

To our knowledge, there are no risk stratification methods to improve referrals for rheumatic and musculoskeletal diseases. In this retrospective study, a language model and conformal prediction-based method have been developed to detect and differentiate inflammatory conditions and non-inflammatory conditions using data routinely available during referrals in primary care. The models were based on routinely available referral data, making them ready for wider validation and amendable to detect other diseases during referrals. We will implement our model in the electronic health record system and referral system to inform about individual diseases and aid referral triage decision-making.

CRediT authorship contribution statement

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

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.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.inffus.2025.103068.

Data availability

The authors do not have permission to share data.

References

[1] G.B.o.D.C. Network, Global Burden of Disease Study 2019 (GBD 2019) Results, Institute for Health Metrics and Evaluation (IHME), Seattle, United States, 2020.
[2] V. arthritis, The State of Musculoskeletal Health 2021, Versus Arthritis (VA), 2021 [Online]. Available: https://www.versusarthritis.org/media/24238/state-of-msk-health-2021.pdf.
[3] M. Al Maini, et al., A global perspective on the challenges and opportunities in learning about rheumatic and musculoskeletal diseases in undergraduate medical education, Clin. Rheumatol. 39 (3) (2020) 627–642, https://doi.org/10.1007/s10067-019-04544-y, 2020/03/01.
[4] A. Cieza, K. Causey, K. Kamenov, S.W. Hanson, S. Chatterji, T. Vos, Global estimates of the need for rehabilitation based on the Global Burden of Disease study 2019: a systematic analysis for the Global Burden of Disease Study 2019, The Lancet 396 (10267) (2020) 2006–2017.
[5] WHO. "Musculoskeletal health." https://www.who.int/news-room/fact-sheets/detail/musculoskeletal-conditions#:~:text=A%20recent%20analysis%20of%20Global,and%20rheumatoid%20arthritis%20(1). (accessed July 30, 2023).

[6] S. Fausto, C. Marina, F. Sonia, C. Alessandro, G. Marwin, The impact of different rheumatic diseases on health-related quality of life: a comparison with a selected sample of healthy individuals using SF-36 questionnaire, EQ-5D and SF-6D utility values, Acta Bio Medica: Atenei Parmensis 89 (4) (2018) 541.

[7] C. f. D. Control and Prevention, Health-related quality of life among adults with arthritis–behavioral risk factor surveillance system, 11 states, 1996-1998, MMWR. Morbid. Mortal. Weekl. Rep. 49 (17) (2000) 366–369.

[8] L. March, et al., Burden of disability due to musculoskeletal (MSK) disorders, Best Practic. Res. Clin. Rheumatol. 28 (3) (2014) 353–366.

[9] M. Cross, et al., The global burden of rheumatoid arthritis: estimates from the global burden of disease 2010 study, Ann. Rheum. Dis. 73 (7) (2014) 1316–1322.

[10] D.G. Hoy, et al., The global burden of musculoskeletal conditions for 2010: an overview of methods, Ann. Rheum. Dis. 73 (6) (2014) 982–989.

[11] E. Savvateeva, O. Smoldovskaya, G. Feyzkhanova, A. Rubina, Multiple biomarker approach for the diagnosis and therapy of rheumatoid arthritis, Crit. Rev. Clin. Lab. Sci. 58 (1) (2021) 17–28.

[12] Z. Chen, et al., ResNet18DNN: prediction approach of drug-induced liver injury by deep neural network with ResNet18, Brief. Bioinform. 23 (1) (2022) bbab503.

[13] Z. Chen, et al., The prediction approach of drug-induced liver injury: response to the issues of reproducible science of artificial intelligence in real-world applications, Brief. Bioinform. 23 (4) (2022) bbac196.

[14] B. Wang, W. Li, A. Bradlow, E. Bazuaye, A.T. Chan, Improving triaging from primary care into secondary care using heterogeneous data-driven hybrid machine learning, Decis. Support. Syst. 166 (2023) 113899.

[15] A. Tiulpin, et al., Multimodal machine learning-based knee osteoarthritis progression prediction from plain radiographs and clinical data, Sci. Rep. 9 (1) (2019) 20038.

[16] L. Folle, et al., Deep learning-based classification of inflammatory arthritis by identification of joint shape patterns—how neural networks can tell us where to "deep dive" clinically, Front. Med. (Lausanne) (2022) 607.

[17] A. Parashar, R. Rishi, Early detection of rheumatoid arthritis in knee using deep learning, in: Proceedings of the International Conference on Data Science, Machine Learning and Artificial Intelligence, 2021, pp. 231–236.

[18] J.K.H. Andersen, et al., Neural networks for automatic scoring of arthritis disease activity on ultrasound images, RMD. Open. 5 (1) (2019) e000891.

[19] L. Folle, et al., Advanced neural networks for classification of MRI in psoriatic arthritis, seronegative, and seropositive rheumatoid arthritis, Rheumatology. 61 (12) (2022) 4945–4951.

[20] Y. Shi, et al., Advancing precision rheumatology: applications of machine learning for rheumatoid arthritis management, Front. Immunol. 15 (2024) 1409555.

[21] L. Bai, Y. Zhang, P. Wang, X. Zhu, J.-W. Xiong, L. Cui, Improved diagnosis of rheumatoid arthritis using an artificial neural network, Sci. Rep. 12 (1) (2022) 9810.

[22] B. Mehta, et al., Machine learning identification of thresholds to discriminate osteoarthritis and rheumatoid arthritis synovial inflammation, Arthritis Res. Ther. 25 (1) (2023) 31.

[23] J. Xiao, R. Wang, X. Cai, Z. Ye, Coupling of co-expression network analysis and machine learning validation unearthed potential key genes involved in rheumatoid arthritis, Front. Genet. 12 (2021) 604714.

[24] D.E. Orange, et al., Identification of three rheumatoid arthritis disease subtypes by machine learning integration of synovial histologic features and RNA sequencing data, Arthrit. Rheumatol. 70 (5) (2018) 690–701.

[25] R. Mitra, et al., Learning from data with structured missingness, Nat. Mach. Intell. 5 (1) (2023) 13–23.

[26] M. Gräf, et al., Comparison of physician and artificial intelligence-based symptom checker diagnostic accuracy, Rheumatol. Int. 42 (12) (2022) 2167–2176.

[27] M. Krusche, J. Callhoff, J. Knitza, N. Ruffer, Diagnostic accuracy of a large language model in rheumatology: comparison of physician and ChatGPT-4, Rheumatol. Int. 44 (2) (2024) 303–306.

[28] T. Savage, A. Nayak, R. Gallo, E. Rangan, J.H. Chen, Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine, NPJ. Digit. Med. 7 (1) (2024) 20.

[29] Q. Zhu, H. Wang, B. Xu, Z. Zhang, W. Shao, D. Zhang, Multimodal triplet attention network for brain disease diagnosis, IEEe Trans. Med. ImAging 41 (12) (2022) 3884–3894.

[30] A. Kline, et al., Multimodal machine learning in precision health: a scoping review, NPJ. Digit. Med. 5 (1) (2022) 171.

[31] L. Liu, et al., A joint multi-modal learning method for early-stage knee osteoarthritis disease classification, Heliyon. 9 (4) (2023).

[32] M.H. Chagahi et al., "Enhancing osteoporosis detection: an explainable multi-modal learning framework with feature fusion and variable clustering," arXiv preprint arXiv:2411.00916, 2024.

[33] Z. Zhou, et al., RATING: medical knowledge-guided rheumatoid arthritis assessment from multimodal ultrasound images via deep learning, Patterns 3 (10) (2022).

[34] R.J. Chen, et al., Algorithmic fairness in artificial intelligence for medicine and healthcare, Nat. Biomed. Eng. 7 (6) (2023) 719–742, https://doi.org/10.1038/s41551-023-01056-8, 2023/06/01.

[35] H. Olsson, et al., Estimating diagnostic uncertainty in artificial intelligence assisted pathology using conformal prediction, Nat. Commun. 13 (1) (2022) 7761, https://doi.org/10.1038/s41467-022-34945-8. /12/15 2022.

[36] N. England. "Advice and guidance." https://www.england.nhs.uk/elective-care-transformation/best-practice-solutions/advice-and-guidance/(accessed 7 May, 2024).

[37] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Adv. Neural Inf. Process. Syst. 30 (2017).

[38] T.K. s. Fund. "Activity in the NHS. The NHS in a nutshell." https://www.kingsfund.org.uk/projects/nhs-in-a-nutshell/NHS-activity#:~:text=In%202022%20there%20were%20an,were%20still%20face%20to%20face (accessed September, 2023).

[39] N. M. B. R. Centre. "Rheumatic and musculoskeletal diseases." https://www.manchesterbrc.nihr.ac.uk/our-research/rheumatic-and-musculoskeletal-diseases/ (accessed May, 2023).

[40] NRAS, The National early inflammatory Arthritis Audit (NEIAA), Natl. Rheumatoid Arthrit. Soc. (2021). Second Annual Report.

[41] G. online. "One in three GP appointments for patients on record NHS waiting list." https://www.gponline.com/onethree-gp-appointments-patients-record-nhs-waiting-list/article/1832759 (accessed September, 2023).

[42] B. Odebiyi, B. Walker, J. Gibson, M. Sutton, S. Spooner, K. Checkland, Eleventh National GP Worklife Survey, University of Manchester: Policy Research Unit in Commissioning and the Healthcare System Manchester Centre for Health Economics, 2021.

[43] J.N. Acosta, G.J. Falcone, P. Rajpurkar, E.J. Topol, Multimodal biomedical AI, Nat. Med. 28 (9) (2022) 1773–1784, https://doi.org/10.1038/s41591-022-01981-2, 2022/09/01.

[44] T. Baltrušaitis, C. Ahuja, L.-P. Morency, Multimodal machine learning: a survey and taxonomy, IEEe Trans. Pattern. Anal. Mach. Intell. 41 (2) (2018) 423–443.

[45] J. Lipkova, et al., Artificial intelligence for multimodal data integration in oncology, Cancer Cell 40 (10) (2022) 1095–1110.

[46] S.-C. Huang, A. Pareek, S. Seyyedi, I. Banerjee, M.P. Lungren, Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines, NPJ. Digit. Med. 3 (1) (2020) 136.

[47] S.R. Stahlschmidt, B. Ulfenborg, J. Synnergren, Multimodal deep learning for biomedical data fusion: a review, Brief. Bioinform. 23 (2) (2022) bbab569.

[48] S. Seoni, V. Jahmunah, M. Salvi, P.D. Barua, F. Molinari, U.R. Acharya, Application of uncertainty quantification to artificial intelligence in healthcare: a review of last decade (2013–2023), Comput. Biol. Med. (2023) 107441.

[49] L. Wells, T. Bednarz, Explainable ai and reinforcement learning—a systematic review of current approaches and trends, Front. Artif. Intell. 4 (2021) 550030.

[50] Z. Chen, et al., Exploring explainable AI features in the vocal biomarkers of lung disease, Comput. Biol. Med. 179 (2024) 108844.

[51] D. Seuß, "Bridging the gap between explainable AI and uncertainty quantification to enhance trustability," arXiv preprint arXiv:2105.11828, 2021.

[52] J. Gawlikowski, et al., A survey of uncertainty in deep neural networks, Artif. Intell. Rev. 56 (Suppl 1) (2023) 1513–1589.

[53] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, "Concrete problems in AI safety," arXiv preprint arXiv:1606.06565, 2016.

[54] C. Leibig, V. Allken, M.S. Ayhan, P. Berens, S. Wahl, Leveraging uncertainty information from deep neural networks for disease detection, Sci. Rep. 7 (1) (2017) 1–14.

[55] Y. Gal, Z. Ghahramani, Dropout as a bayesian approximation: representing model uncertainty in deep learning. International Conference on Machine Learning, PMLR, 2016, pp. 1050–1059.

[56] A.N. Angelopoulos, S. Bates, A gentle introduction to conformal prediction and distribution-free uncertainty quantification, arXiv preprint arXiv:2107.07511, 2021.

[57] L. Mossina, J. Dalmau, L. Andéol, Conformal semantic image segmentation: post-hoc quantification of predictive uncertainty, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 3574–3584.

[58] "Indices of deprivation." https://ckan.publishing.service.gov.uk/dataset/indices-of-deprivation1 (accessed 13 June, 2024).

[59] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: 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, Association for Computational Linguistics, 2019, pp. 4171–4186, jun.

[60] 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, Florence, Italy Association for Computational Linguistics, 2019, pp. 3651–3657, https://doi.org/10.18653/v1/P19-1356.

[61] A. Vaswani, et al., Attention is all you need, in: Advances in Neural Information Processing Systems, 30, Curran Associates, Inc., 2017, pp. 5998–6008.

[62] M. Ganaie and M. Hu, "Ensemble deep learning: A review," arXiv preprint arXiv:2104.02395, 2021.

[63] Y. Feng, X. Wang, J. Zhang, A heterogeneous ensemble learning method for neuroblastoma survival prediction, IEEe J. Biomed. Health Inform. (2021).

[64] A. Angelopoulos, S. Bates, J. Malik, M.I. Jordan, Uncertainty Sets for Image Classifiers using Conformal Prediction, ICLR, 2021.

[65] V. Vovk, A. Gammerman, G. Shafer, Algorithmic Learning in a Random World, Springer, 2005.

[66] S.M. Lauritsen, et al., Explainable artificial intelligence model to predict acute critical illness from electronic health records, Nat. Commun. 11 (1) (2020) 3852, https://doi.org/10.1038/s41467-020-17431-x, 2020/07/31.

[67] N. Rodis, C. Sardianos, P. Radoglou-Grammatikis, P. Sarigiannidis, I. Varlamis, G. T. Papadopoulos, Multimodal explainable artificial intelligence: A comprehensive review of methodological advances and future research directions, IEEe Access. (2024).

[68] E.J. Topol, High-performance medicine: the convergence of human and artificial intelligence, Nat. Med. 25 (1) (2019) 44–56.

[69] L. Parcalabescu and A. Frank, "Mm-shap: a performance-agnostic metric for measuring multimodal contributions in vision and language models & tasks," arXiv preprint arXiv:2212.08158, 2022.

[70] NRAS, The National Early Inflammatory Arthritis Audit (NEIAA), National Rheumatoid Arthritis Society, 2023. Fourth Annual Report.
[71] D. Vela, A. Sharp, R. Zhang, T. Nguyen, A. Hoang, O.S. Pianykh, Temporal quality degradation in AI models, Sci. Rep. 12 (1) (2022) 11654.
[72] L. Liou, et al., Assessing calibration and bias of a deployed machine learning malnutrition prediction model within a large healthcare system, NPJ. Digit. Med. 7 (1) (2024) 149.