UNDERSTANDING FAIRNESS AND EXPLAINABILITY IN MULTIMODAL APPROACHES WITHIN HEALTHCARE

Sophie Martin
UCL
Department of Computer Science
s.martin.20@ucl.ac.uk

Jonathan Pearson
NHS England
Data Science, Transformation Directorate
jonathanpearson@nhs.net

ABSTRACT

Explainability, fairness, and bias identification and mitigation are all essential components for the integration of artificial intelligence (AI) solutions in high-stake decision making processes such as in healthcare. Whilst there have been developments in strategies to generate explanations and various fairness criteria across models, there is a need to better understand how multimodal methods impact these behaviours. Multimodal AI (MMAI) provides opportunities to improve performance and gain insights by modelling correlations and representations of data of different types. These approaches are incredibly powerful for the analysis of healthcare data, where the integration of data sources is key for gaining a holistic view of individual patients (personalised medicine) or evaluating models across different patient profiles to ensure safe and ethical use (population health). However, MMAI presents an unique challenges when deciding how best to incorporate and fuse information, maintaining an understanding of how data is processed (explainability), and ensuring bias is not amplified as a result. Here, we explore a case study, **Multimodal Fusion of Electronic Health Data for Length-of-Stay Prediction**, with a focus on treating time-series and static electronic health data as distinct modalities. We evaluate and compare different methods for fusing data in terms of predictive performance and various fairness metrics. Additionally, we apply SHAP to highlight the influence of specific features and explore how such explanations can be used to reveal or confirm bias in the underlying data. Our results showcase the importance of modelling time-series data, and an overall robustness to bias compared to unimodal approaches across various fairness metrics. We also describe exploratory analysis which can be conducted and developed further to mitigate bias post-hoc, or gain further insights into the relative importance of specific modalities from multimodal models. This work was conducted as part of an NHS England PhD Internship.

Keywords: Fairness, bias, explainable AI, multimodal, electronic health records, fusion

Code and documentation: www.github.com/nhsengland/mm-healthfair.

Contents

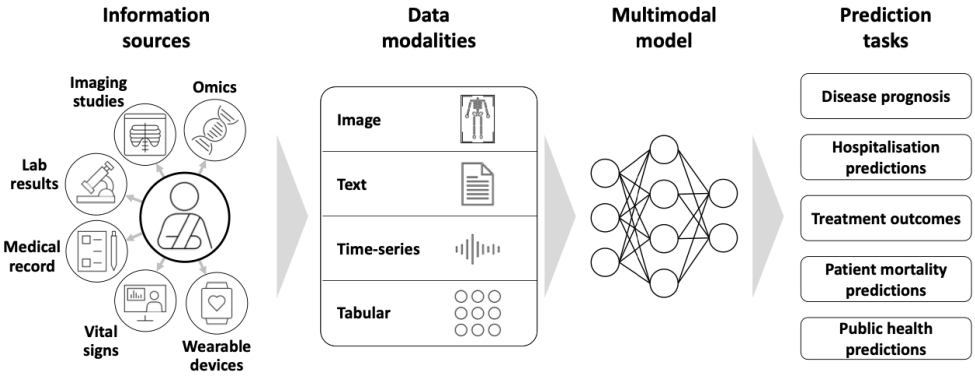


Figure 1: Schematic diagram of modalities associated with healthcare data that can be used for a variety of tasks. Taken from [29].

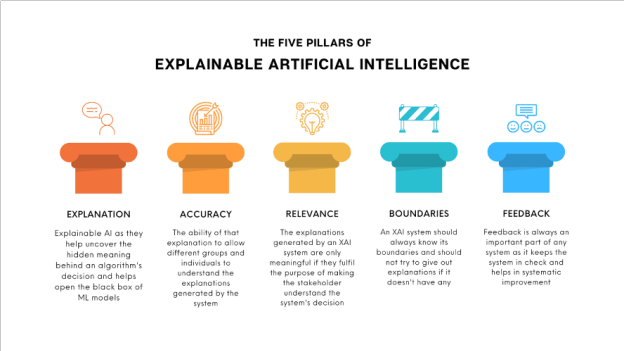
ICU Stay ID	25130					
Time-Invariant Data	*					
Time-Series Data (Discrete Events)	** *	*		** **		* *
Time-Series Data (Continuous Events)						
Clinical Notes			*			*
Hours	1	2	3	4	5	6

Figure 2: Electronic health data can be considered multi-modal due to the presence of time-varying (discrete and continuous), static data (demographics) and clinical notes, taken from [57].

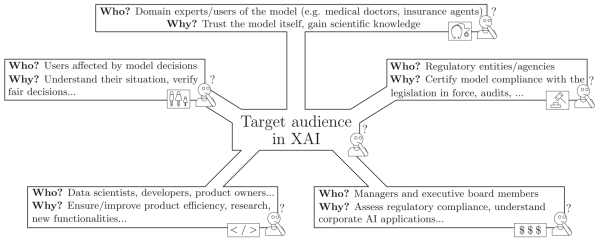
1 Introduction

Multimodal approaches have been explored across a variety of healthcare tasks. Figure 1 is taken from a review of multimodal AI tasks across healthcare and summarises the differences sources of patient information that can be obtained and combined in a multimodal model for different downstream tasks. Recent studies have focused largely on the integration of imaging and text data with additional modalities, in an effort to leverage the advancements of deep neural networks and large language models (LLMs). A review by Pei and colleagues [40] highlighted that many multimodal studies focus on the vision-language paradigm. Whilst this is an exciting area of research in healthcare, there is also opportunity in exploring the integration of more readily available data types such as time-series and clinical notes from routinely-collected electronic health records (EHR). For example, Yang and colleagues [57] highlight the multi-modal nature of EHR data due to the presence of continuous and discrete time-varing data which is inherently different in structure to tabular static information, shown in Figure 2. The availability of large scale open EHR datasets can foster faster development of modelling approaches whilst avoiding some of the drawbacks of requiring access to sensitive data such as omics, imaging and text. Moreover, learning from these sources and approaches can then be applied to different contexts and help to guide research on more sensitive modalities.

Explainable AI (XAI) is used to describe a set of techniques that aim to provide human-understandable explanations for the decisions generated by AI models. This is useful not only for model developers, who can use XAI to debug models and identify potential biases, but it also plays a role in creating more trustworthy and transparent systems for regulators and end-users. Explainability has also attracted attention in recent years as a potential component of developing trustworthy systems with outputs that can be understood, interpreted and justified. Not only does this help to increase trust in the model, but it is a necessary requirement to fulfill UK and EU GDPR laws, which include a "right to an explaination" for potential stakeholders [47]. Figure 3a highlights 5 key pillars of explainable AI: explanation,



(a) The five pillars of explainable AI, as described in [8].



(b) The different target audiences and stakeholders of explainable AI, from [10].

Figure 3: Pillars of explainable AI and overview of different target audiences.

accuracy, relevance, boundaries and feedback. These summarise the different aspects that must be balanced when developing and evaluating XAI methods. Attention has been heavily placed on explanation and accuracy, however relevance, boundaries and feedback are vital, particularly when considering the use of XAI for healthcare AI systems, and demonstrate its utility for AI transparency, ethics and regulation.

Many different techniques for explaining model predictions have been proposed from recent research. However, it is not trivial to ascertain which method is most appropriate for a given model, data, or task. Often this decision depends on the target audience, as there are different perspectives on the utility of XAI depending on the intended application, as shown in Figure 3b. As the field of XAI continues to grow, and efforts look towards formalism, regularised frameworks for XAI validation and introduction of ideas around causality for deriving more meaningful and actionable insights [46]. These challenges all become even more relevant in multimodal approaches where the goal is not only to understand how features of a given modality are processed and used, but also to disentangle to relationships between modalities in a clear, justifiable way. Moreover, understanding the role of XAI in healthcare applications presents a unique challenge, partly due to the heterogeneous, unstructured and large volume of associated data. There are also nuances around defining useful metrics that accurately reflect the risks and impact of model decisions on healthcare outcomes. Therefore applying techniques to healthcare datasets often requires expert knowledge to guide insights and ensure the utility of such systems can be exploited, quantified and measured against any potentially adverse effects.

The importance of fairness and bias in AI has gained the recognition of researchers, policymakers, and industry. This has been partly driven by studies that have identified biases against certain groups in existing systems, such as facial recognition, judicial decision making and hiring processes [18]. Biased systems are dangerous as they can perpetuate systemic discrimination and inequality, with detrimental effects on individuals and subpopulations. Fairness is key component of ethical AI frameworks and has attracted attention as companies and policymakers consider how to regulate and ensure that AI models are safe before large-scale deployment. One approach to mitigating against bias is to ensure models are fair by quantifying their performance within subpopulations or against certain sensitive attributes. This can reveal whether biases, typically present in the underlying data, manifest in differences in the behaviour of the AI model [16]. Bias can also arise from decisions made along the development processes such as assumptions inherent to the algorithm (model bias) or injected by the user consciously or unconsciously such as during model selection based on desired performance criteria (user bias). Multimodal modelling can then present an opportunity to explore whether bias is exacerbated or inhibited as a result of combining data from multiple sources or forms. Moreover, some modalities may be more susceptible to types of bias than others, and understanding this interplay could be useful for informing decisions around data collection, processing and sharing.

2 Background

2.1 Multimodal AI

Multimodal AI is a rapidly developing field that leverages different data modalities, such imaging, text, time-series and omics, to create comprehensive and accurate models. By combining different types of data from various sources, multimodal AI can capture complex relationships and patterns that might be missed by considering a single modality alone. Liang propose six challenges that frame the different aspects of Multimodal AI [32]: representation, alignment, reasoning, generation, transference, and quantification. This taxonomy reflects the flow and process of tackling multimodal modelling, which typically start out with representation and in some cases alignment. Representation and

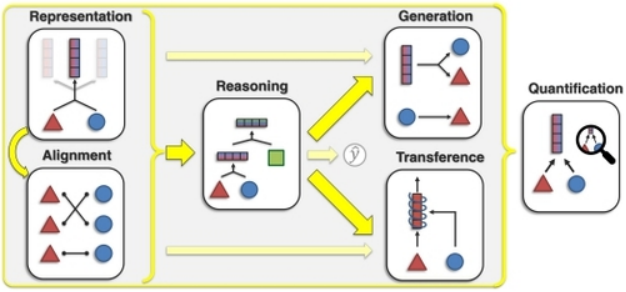


Figure 4: The six challenges of multimodal AI as proposed by [32].

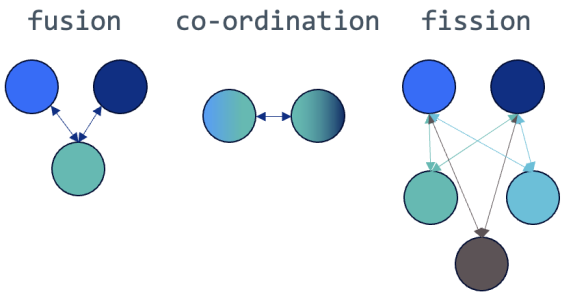


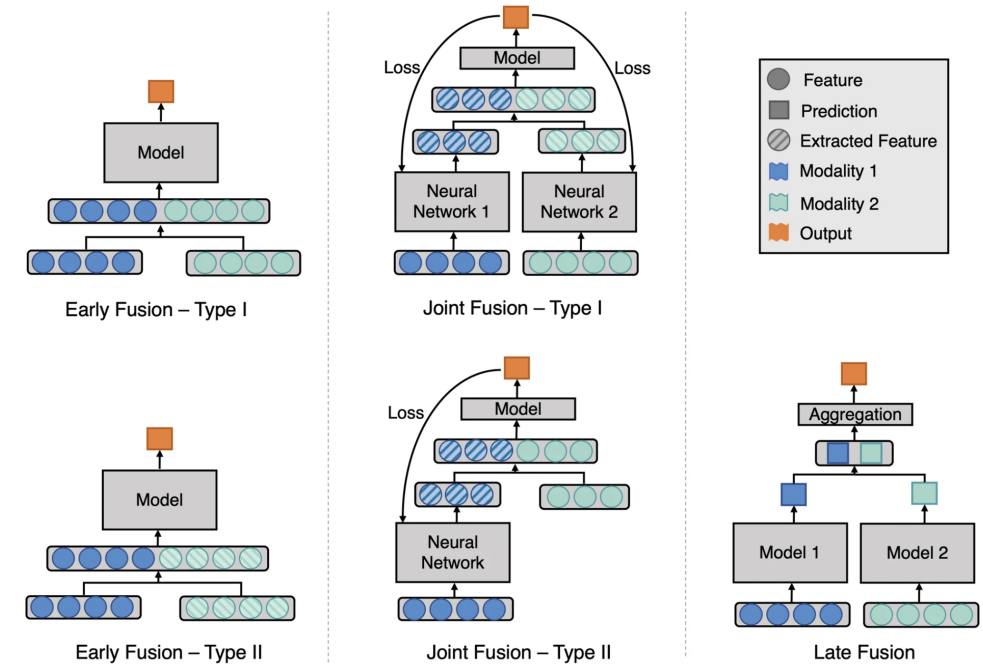
Figure 5: Schematic overview of multimodal approaches inspired by [32].

alignment are the initial step in the pipeline and focus on the both heterogeneity and within-modality connections as well as interactions between modalities. Once this has been obtained, reasoning aims to combine the information in a way that is specific and useful for the problem. Although most applications involve multimodal modelling for prediction, other applications can include generation, where additional modalities are formed from learned cross-modal interactions. On the other hand, transference describes the transfer of information from high-resource modalities to low-resource ones and their representations. Finally, quantification refers to obtaining a deeper insights of the different modalities and their interactions.

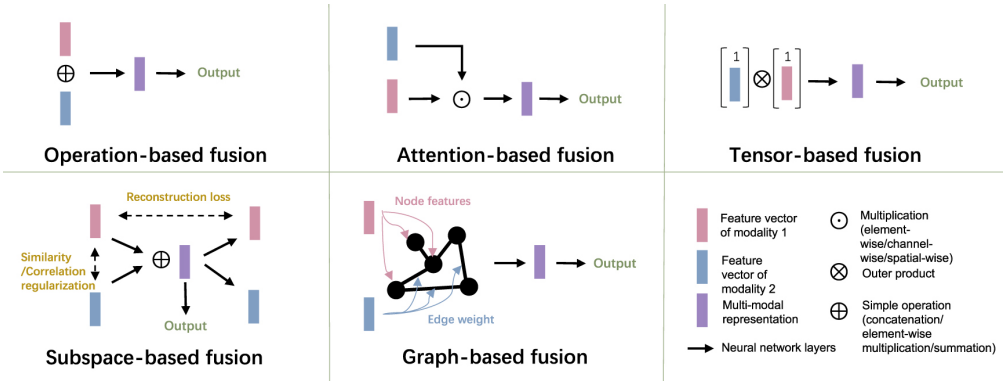
Representation is the first step towards utilising multimodal data, and is the focus of the work in this project. Within that, there are many ways to represent information across modalities which often depend on the use case, data structure and context. These different modelling approaches can be categorised into three types: fusion, co-ordination and fission as shown in Fig 5. These can be differentiated by the number of input modalities compared to the number of representations [32]. Fusion methods are most popular: where data is combined, often into a shared embedding space through learned representations. In this case, the number of input modalities is greather than the number of representations learned $N_{mod} > N_{rep}$. The latent space can then be sampled to produce predictions. For example, Aksoy et al. [6] present a framework for generating chest x-ray reports from structured EHR data (e.g., vital signs), x-rays, and clinical notes. Co-ordination considers how information from additional modalities can help inform another or provide additional context. In this case, the number of input modalities equals the number of learned representations, $N_{mod} = N_{rep}$. Co-ordination differs from fusion since rather than trying to create an alignment representation space from the input modalities, the emphasis lies in finding useful information from each respective modality. This often involves some measure of similarity such as cosine similarity, or applying a contrastive loss, which encourages agreement between positive pairs and maximises disagreement between negative pairs in the embedding space. Co-ordination is often beneficial when the data modalities are very different in their structure such as image and text. Fission models are the more complex approach, where separate models can be produced based on different combinations or representations of the data and $N_{mod} < N_{rep}$. This broader set of representations can reflect different aspects about the knowledge shared across modalities such as details about the data structure from clustering.

2.1.1 Fusion

We focus on fusion methods, which can be further categorised depending on the time at which modalities are combined i.e., early, joint or late. These are depicted in Figure 6a. Some studies have explored whether one type is better than



(a) Schematic representation of early, joint and late fusion from [26].



(b) Summary of fusion methods available in `fusilli` package [52].

Figure 6: Overview of different stages and types of fusion.

the other. For example, [23] Hayet and colleagues fuse clinical time-series data with chest x-rays and explore early, joint and late fusion methods for in-hospital mortality prediction and phenotypic classification. They present a novel approach MedFuse as an alternative to vanilla early and joint fusion, and report some advantage of their model over them. However, overall there was little difference found between early and joint fusion. Similarly, Huang et al., [26] explore a variety of fusion types for combining imaging with EHR data, and report that early fusion was most promising. Additionally, fusion methods can also be grouped by the operations used as seen in 6b. Deciding on the best approach is usually an experimental process, aided by toolkits such as `fusilli` [52][1] which currently supports the evaluation of different methods for tabular-tabular or imaging-tabular fusion. Similarly, Lawry-Aguila and colleagues provide a package for exploring different autoencoder architectures for learning joint representation from multimodal data called `multi-view-AE` [4][2].

[1] `https://fusilli.readthedocs.io/en/latest/`
[2] `https://github.com/alawryaguila/multi-view-AE`

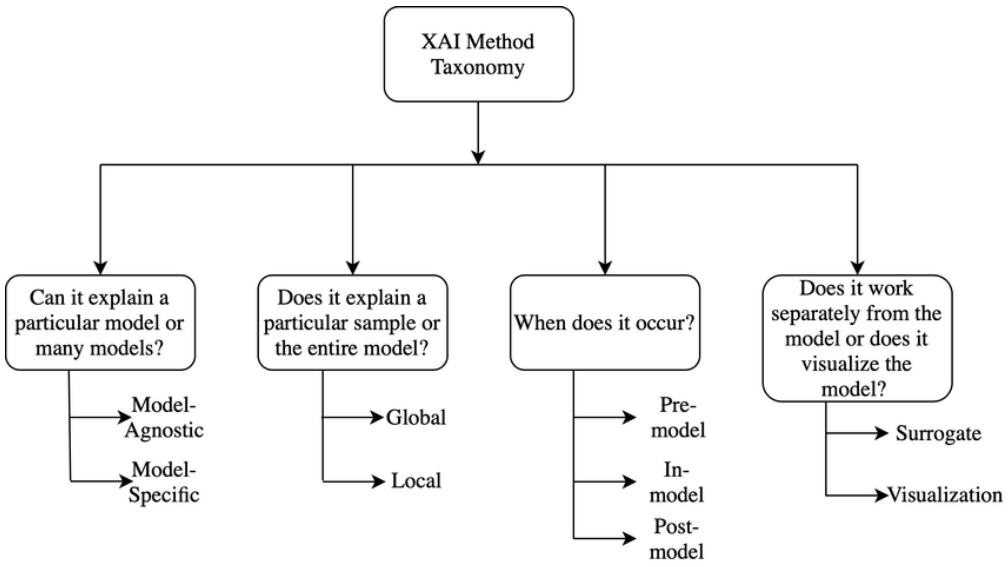


Figure 7: Example taxonomy for categorising XAI approaches.

2.2 Explainable AI

Explainable AI is used to describe a set of techniques that aim to shed light on the reasoning behind an AI model's predictions. For a through introduction to the different techniques, Christoph Molnar's online e-book provides a tangible and accessible overview of the most popular methods [38]. Whilst there have been several different taxonomies and frameworks proposed over recent years, a common framework for categorising these techniques is shown in Figure 7. These categorisations are useful for understanding the limitations of a given approach, and are important to consider when interpreting the results of any given method. The first categorisation refers to the generalisability of the method. Techniques such as LIME [44] and SHAP [34] which are becoming popular tools in several domains, are considered model-agnostic, as they can be applied across many different machine learning algorithms or deep learning architectures. On the other hand, model-specific techniques such as Grad-CAM [48] or layer-wise relevance propagation [13] have been designed for the explanation of specific types of models such as convolutional neural networks.

2.2.1 The Taxonomy of Explainable AI

Moreover, global methods describe techniques which provide an insight into the workings of the model, typically by considering all samples seen during training. Local explanations are used to provide instance-wise outputs, which is often more desirable when applying XAI in healthcare where patient-level specificity could play a role in moving towards personalised medicine. Similarly, considering at which point the explanation is generated helps to distinguish between post-hoc or post-model methods and the development of inherently interpretable algorithms (in-model). This differentiation mirrors the debate around interpretability versus explainability where the latter is typically used to refer to in-model or humanly-interpretable algorithms such as decision trees. In contrast, post-model methods are applied after model training (explainability) and are often used for black-box models or deep neural networks where the lack of transparency arises largely due to the large number of parameters and complexity of the network. Finally, surrogate explanations describe a set of methods which produce secondary models that aim to capture the key features of a learned model's decision space. Often these surrogate models are much less complex, or interpretable and can be used to summarise what the the initial model has learned. Visualisation methods may be most applicable to XAI for graphical neural networks, offering a visual representation of the salient connections.

2.2.2 The Challenges of Explainable AI

Some challenges and implications of explainable AI are thoroughly discussed in a widely-cited review paper by Barredo Arrieta and colleagues [10]. One of these is the difficulty in validating model explanations, and balancing the use of XAI for knowledge discovery against confirming existing knowledge. This dilemma is also captured in an article by Bienefeld and colleagues [12] (see Figure 8). Here, the emphasis lies on the use of explainable AI to listen to the data (ML developer mental model) and see whether information about bias can be revealed.

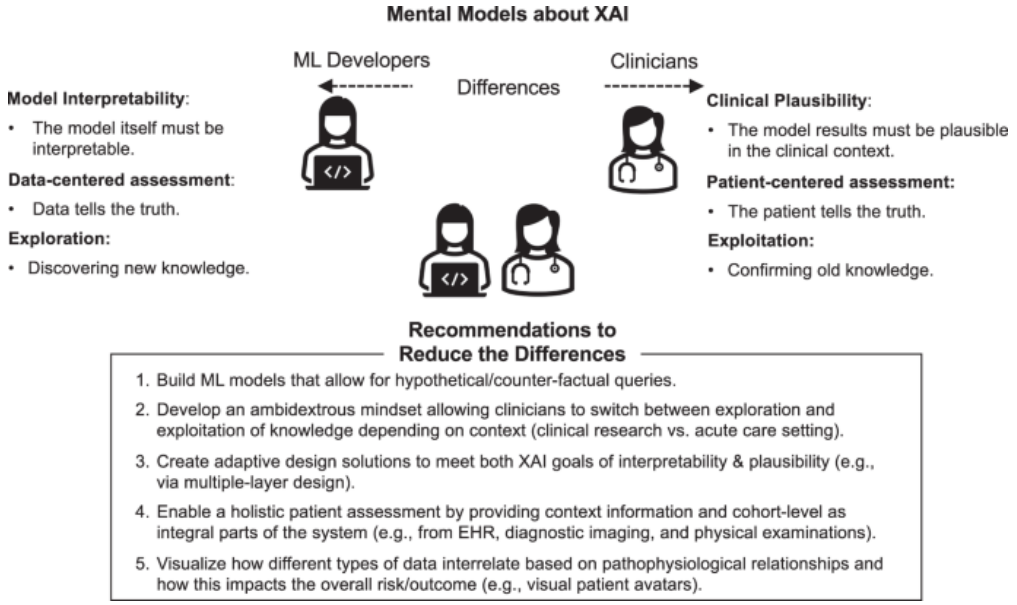


Figure 8: The different needs of developers and clinicians for Explainable AI.

Another challenge lies in how to apply Explainable AI to multimodal models. The fusion of information from different data sources and structures means that it is often not-trivial to apply uni-modal methods to multimodal models. Model agnostic post-hoc explanation toolboxes have been used for some multimodal studies as they are flexible and can reveal insights as to which data modality is most useful [49, 42]. An initial approach could be to consider specific data types, and use cases individually, to highlight the advantages and disadvantages of different techniques before attempting to consider multiple streams. However, it may be desirable to produce modality-specific explanations not only disentangle the most important data type, e.g., saliency mapping within an image, or identification of important words in a sentence, within a multimodal context. In these cases, more complex approaches must be taken to overcome the challenge of handling data types of varying dimensionality. Moreover, the use of XAI in healthcare has been debated. For example, Ghassemi and colleagues [20], highlight the limitations and example failure cases of explanation techniques. Additionally a survey amongst NHS healthcare workers revealed that whilst many have a desire for explainability, there is uncertainty around its suitability for individual case-level analysis, due to many of the limitations of current techniques described above [1].

2.3 Bias and Fairness

For the safe and ethical deployment of AI systems, it is imperative that we are able to quantify and demonstrate that they can be used fairly. Fairness in AI describes a set of criteria and associated guidelines to encourage safety, minimise bias and ensure ethical requirements are met. As such, rather than a single measure, fairness in AI encompasses a broad range of standards and frameworks, which can be specific to certain region, businesses or application depending on their own legal and ethical definitions.

2.3.1 Types of Bias

Broadly, there are three channels in which unfair or harmful practices may enter a AI system: from the user to the data, the data to the algorithm, or the algorithm to the user (and also their reverse). As such, by considering the end-to-end life cycle of the an AI system from design to development to deployment, one can highlight various touch points and questions to ascertain how fairness and bias need to be considered and addressed. This is nicely visualised in Figure 9 taken from this blog post [11]. The three areas reflect the different ways bias can enter or be introduced into an AI system, consciously or subconsciously. For example, bias can be introduced during the design process such as as a result of over or undersampling data from a specific group. Bias can also arise during development (data to algorithm) such as user-defined optimization criteria which can unfairly favour a certain group, or choices around training data splitting, feature optimisation and label definitions. These can all embed bias into the model itself, and exacerbate existing imbalance. Finally, bias from the model to the user reflects bias during deployment. This can refer to metrics used for evaluation which may infer production-level decisions, or bias in the model outputs themselves which can lead

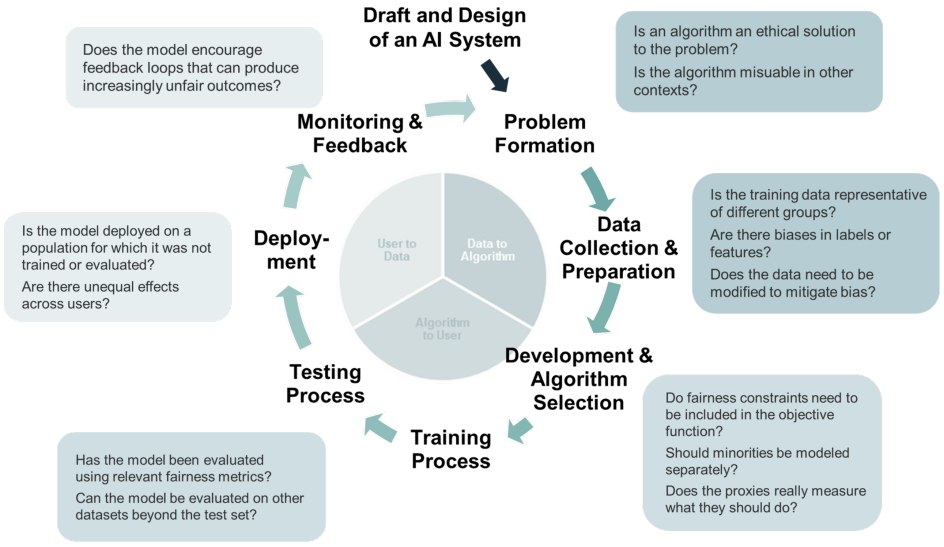


Figure 9: Example questions around bias and fairness along the AI design and development cycle. Figure taken from [11].

to feedback loops and increase unfairness. As a result of these channels, there are different types of bias that occur. A blog by data development platform Encord [17], list six different types of bias, although these are not necessarily exhaustive. These are listed and categorised in Table 1.

2.3.2 Mitigating Bias

AI Blindspot [3], devised by a team at MIT [5], is a framework for characterising the different ways that developers can be 'blind' to bias in AI, which can be useful when designing a new study or evaluating existing pipelines. This is depicted in Figure 10. The three key sources of bias are mirrored in the different ways bias can be mitigated, typically categorised by from an developers perspective as either: before, during or after model training. These have been outlined in a blog post by Holistic AI [25] and are summarised as follows:

Strategies for mitigating bias prior to training often rely on shifting the underlying data distribution. For example, many studies use sampling techniques such as the Synthetic Minority Oversampling Technique (SMOTE) to address class imbalance, and produce augmented samples of the minority class [15]. Alternatively, an additional feature engineering step can be applied to learn fairer representations of the data as proposed by Zemel and colleagues [59]. This procedure involves finding a latent representation which encodes the data, minimising the information loss from non-protected attributes, whilst removing information about protected attributes.

Moreover, some techniques can be applied during the algorithm learning stage. These typically involve adjusting or influencing the model's loss function and optimisation process. For example, Kamishima1 and colleagues proposed a regularisation term to produce fairness-aware classifiers and remove prejudice [27]. In contrast, adversarial algorithms can be applied, where a second learning scheme is used to correctly make a prediction whilst being blinded to a protected variable by using equality constraints.

There are also mitigation strategies that are applied in a post-hoc fashion. For instance, testing data can be altered, such as via the Gradient Feature Auditing (GFA) method proposed by Adler and colleagues [3], which evaluates the influence of each feature on the trained model and adjusts them accordingly. Alternatively, learned predictor outcomes can also be adjusted to minimise bias effects. Calibrated Equalized Odds [41] involves two binary classifiers for the privileged and unprivileged groups respectively, such that output probabilities can be adjusted in favour of an equalized odds objective.

[3] `https://aiblindspot.media.mit.edu`

Bias	Channel	Definition	Mitigation(s)
Measurement	User→Data	Systematic errors during data recording e.g., calibration errors, human subjectivity	Statistical adjustments
Omitted Variable	User→Data	Correlated or confounding variables are left out of analysis causing spurious results	Ensuring all relevant variables and interactions are captured
Aggregation	User→Data	Data aggregation which masks underlying patterns	Balance granularity and clarity
Sampling	User→Data	Data is not properly sampled at random introducing inaccurate sources of bias	Careful data analysis, data recollection, resampling techniques
Linking	User→Data	Inaccurate, assumed correlations between variables	Causal analysis and empirical evidence
Labelling	User→Data	Incorrect or biased labels	Multiple raters to reduce subjectivity
Algorithmic	User→ Algorithm	Bias arising from the decision-making process of the algorithm itself	Careful choice of algorithm with fairness in mind
User interaction	Data→ Algorithm	Human bias can be reinforced through feedback or model interactions	Considered human-computer interaction interfaces and reinforcement feedback loops
Popularity	User→ Algorithm	Choice that favour the most popular approach over what is best	Empirical, data-driven algorithm selection
Emergent	Algorithm→ User	Biases learned by the AI even if the underlying data is unbiased	Bias identification
Evaluation	User→ Algorithm	Using biased criteria to measure the performance of models	Selecting metrics with fairness and bias in mind
Historical	Data→ Algorithm	Embedded social or cultural inequalities present in data that can be amplified by AI	Awareness, revised and continual data collection processes
Population	Data→ Algorithm	Over- or under-representation and attention of the model due to data imbalance	Targeted and stratified data collection
Social	Algorithm→ User	Biased predictions arising from cultural attitudes and prejudices	Efforts to remove such prejudices prior to training
Temporal	Algorithm→ User	Predictions are only true for a certain time	Constant, regular model evaluation and monitoring

Table 1: Examples types of bias, their channels and example mitigation. Inspired and adapted from [17].

2.3.3 Fairness metrics

In order to mitigate against bias, it is useful to be able to quantify the precense of bias. As such, another area of interest lies in developing useful metrics that can capture these effects along the ML development and deployment pipeline. Many of these metrics are discussed in this review by Castelnovo and colleagues [14].

For example, demographic parity is a metric that ensures that a models' prediction, $h(X)$, is independent of protected characteristics, e.g., feature A. In a binary classification setting, this translates to equal selection rates across subgroups. Mathematically demographic parity requires that:

$$\mathbb{E}[h(X)|(A = a)] = h(X)\forall a$$

. Whilst this metric is a simple and intuitive measure of fairness across subsets with different characteristics, it can quickly become problematic or unsuitable for certain problems. For example, if some participants are members of

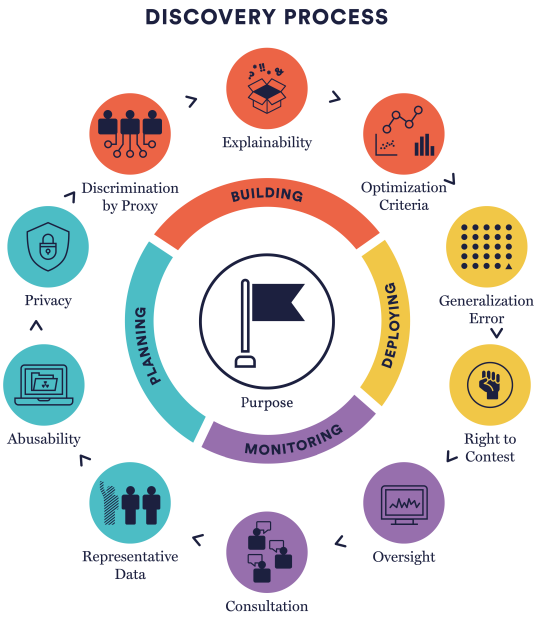


Figure 10: Schematic diagram of different blindspots which often lead to bias [5].

multiple protected characteristics then demographic parity can become unfair on an individual level, as equality cannot be maintained across all subgroups.

As an alternative, equalised odds was proposed to mitigate some of the drawbacks of demographic parity. Equalised odds means that the true positive rates and false positive rates are equal across groups such that:

$$\mathbb{P}(h(X) = 1|Y = 1) = \mathbb{P}(h(X) = 1|Y = 0)\forall a$$

where Y is the true label (binary case). Similarly, equalised opportunity is a relaxed version which only considers the positive class, such that $\mathbb{P}(h(X) = 1|Y = 1)$ is consistent across all groups.

One of the key challenges lies in the fact that often, fairness metrics cannot be fulfilled simultaneously, meaning that a choice must be made as to which to prioritise. This is depicted in Figure 11, which provides a visual example for where equalised opportunity can be fulfilled but not demographic parity.

2.4 Application to Healthcare

Many of these concepts are universal and relevant across different machine learning tasks. However in healthcare, certain decisions must be made to account for the high-stake, and sensitive nature of predictive tasks and complex nature of the underlying data. For example, deciding which multimodal fusion method to choose is non-trivial. Depending on the data sources in question, healthcare datasets are likely to contain data with varying levels of dimensionality, and may contain relevant structure containing useful information. It is also crucial to consider the many dependencies and relationships present in patient data, both between modalities such as an x-ray and radiology report, but also within modalities such as electronic health data where features are highly correlated. Ensuring that the modelling process is able to appropriately capture and handle correlations in data is key in healthcare.

Moreover, explainability is of utmost importance in healthcare settings due to the high-stakes and potential implications on patient care and services. Beyond this, it also key to consider who the explanation is for since this can influence which method is most applicable. For example, group level explanations can be useful for gaining a broad understanding of model behaviour. Developers may be interested in using this type of explanation to identify biases present in the model by examining and comparing explanations for different sub-populations. On the other hand, individual-level explanations may play a crucial part in personalised medicine, by revealing patient-specific insights into which aspects of their data are most relevant. This area of explainable AI is actively developing however, and due to a lack of trust in many XAI outputs, some healthworkers have expressed concern about the overpromise of individual-level XAI in a recent NHS report [1].

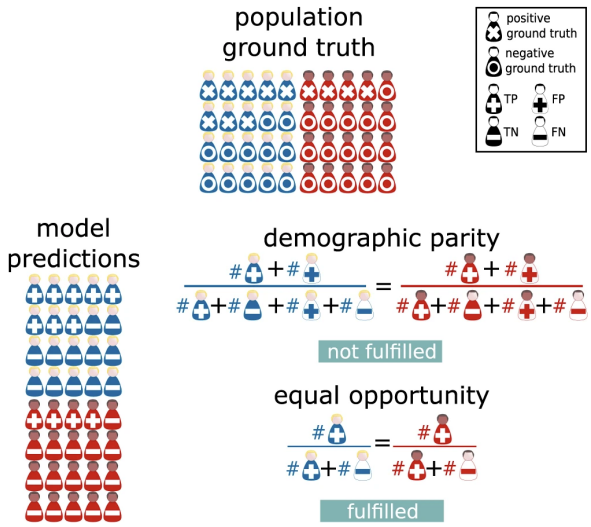


Figure 11: Visual example where fairness metrics cannot be simulaneously fulfilled. Taken from [45].

Additionally, certain fairness metrics may be more or less appropriate when applying these tools to healthcare tasks. For instance, if used in a diagnostic setting, then ensuring equalised odds across different characteristics is essential, to ensure both true positives and false positives are balanced. This is crucial since false positives can also have negative consequences such as influencing treatment decisions and minimising unnecessary invasive procedures.

3 Multimodal Fusion of Electronic Health Data for Length-of-Stay Prediction

The aim of this project is to explore fairness and explainability in the context of multimodal approaches for healthcare. This involves investigating different combinations of modalities present in an open healthcare dataset and evaluating different approaches to multimodal modelling in terms of predictive power, explainability and fairness. To do this, we have utilised multi-modal data from the MIMIC-IV (2018) open dataset which contains hospital and intensive-care unit data, chest x-rays and associated clinical report notes for use in downstream analysis. Here, we focus on a case study which explores the combination of time-series data (lab events, vital signs) with static information (patient demographics) in predicting length-of-stay as a binary classification task (long vs. short). This case study allows us to delve into the impact of modelling choices on downstream evaluation metrics, with a focus on the presence of bias and impact on fairness with respect to protected characteristics. Due to the use of time-series data, we investigate how missingness not-at-random (NAR) can feed into investigations around bias and whether this can be mitigated during or after modelling. This project addresses the following research questions:

1. How does the incorporation of time-dependant information impact model performance?
2. What is the impact of multimodal fusion choices on performance, explainability and fairness?
3. How can explainability be used to identify model bias arising from underlying data?

The project was structured around the following four key milestones:

1. **Data Preparation** The MIMIC database is a very large source of heterogeneous, raw clinical data. Therefore, time will be required to design a suitable data preprocessing and loading pipeline that extracts the relevant variables in a format that can be trained on with the desired models.
2. **Multimodal Modelling (time-series + static)** Here, we treat time-series data (vital signs, laboratory events) and static data (demographics) as distinct modalities due to their different structures, and explore fusion strategies to combine them.
3. **Model explainability and bias** The next milestone will seek to explore different mechanisms for explaining the predictions of the above models such as identifying key features or exploring the interactions between the two modalities. Producing explanations will also provide the opportunity to will involve the evaluate the model fairness. We will carry out several ablation experiments using stratified subsets of the dataset to create algorithmic-biased models e.g., predicting on a subset with label imbalance towards certain demographics.

This will allow a further investigation into whether explanations can help to identify this bias and how this links to the underlying fusion method.

4. **Report and codebase** The final milestone focused on writing these findings in a report and publishing the code used to an open-source GitHub repository following RAP guidelines to enable future work.

3.1 Related Work

This project relates to several studies exploring fairness and bias within the MIMIC dataset as well as those investigating different multimodal approaches. For example, Wang and colleagues conducted a comparison of early, joint and late fusion methods for a range of predictive tasks using MIMIC-III [55]. Specifically, when considering only time-series (time-varying features) with static tabular data, multimodal fusion methods were found to be important for appropriately utilising the temporal information. This has since been explored further by Ma and colleagues [35] who utilised a specific attention gate for cross-modal fusion. Moreover, for fairness and explainability, this study by Meng and colleagues includes a throughout comparison of different deep learning algorithms on a set of extracted features for MIMIC-III [36]. They evaluate a range of interpretability methods such as SHAP, and explore the relationship between feature importance and known protected characteristics such as age to assess algorithmic bias. Inspired by these findings, we focus on the early fusion of time-series and tabular data in this work when applied to MIMIC-IV and evaluated with multiple fairness and explainability methods.

3.2 Data Curation

To investigate the impact of modelling choices on fairness and bias, we aimed to curate a multimodal dataset from MIMIC's EHR data, with a focus on time-series and static data as independent modalities. Whilst, existing pipelines have been made publicly available to encourage reproducible analysis of this popular dataset, however many are tailored towards MIMIC-III version. The advantage of using MIMIC-IV is not only a larger, better structured cohort but also access to separated ED data, obtained prior to a subject's admission into the hospital. In contrast, MIMIC-III only contains data from a hospital and ICU ward stay, the latter of which we consider out-of-scope of this project (see Appendix). To align with reproducible aims of previous work we adapt the benchmark data pipeline[4] produced by Harutyunyan and colleagues on MIMIC-III, to MIMIC-IV [22]. The data curation pipeline is discussed below, including several key decisions made around handling different recording frequencies, feature encoding and handling missingness. A sample stay processed dictionary object is shown in Figure 12.

3.2.1 Extracting the data

To start, we extracted static and time-series data from the MIMIC-IV database: `extract_data.py`. This script reads and processes hospital (HOSP) and emergency department (ED) admissions time-series data, demographic data for each subject stay. Throughout this work we define a stay as an admission to the hospital from the emergency department. We aim to predict the duration time spent in the hospital (length-of-stay), t, such as $t = 0$ refers to the time of hospital admission. Our model utilises data recorded during the ED stay and the first t_{max} hours spent in the hospital. The results in this report utilise $t_{max} = 48$ hours. Static data includes information about the patient such as age, gender, insurance type, marital status and ethnicity. Time-series data was extracted from two events tables: vital signs recorded during the ED stay and also hospital laboratory (lab) events. To speed up reading of lab events data, we also apply a filter to extract specific items. In this work, we restrict to the top 20 items with the most non-zero values (least missingness). Other event tables available in the MIMIC database were considered out-of-scope but could be added to the codebase in future. During this step, we also filtered stays using the following criteria:

- Only include one ED->HOSP stay per patient
- Exclude patients with an age < 18
- Remove any hospital stays without a corresponding ED stay and transfer

3.2.2 Preparing the data

Once the relevant events and stay-level data were collected, we prepared and processed the output files to create a single dictionary object containing all stays (keys) and time-series and static information (values): `prepare_data.py`. This allowed us to clean and filter the data such as removing outliers and filling empty string values. The stays are also split into training and test sets for model development and training using stratification based on the desired length-of-stay duration: `create_train_test.py`. In this work we chose to focus on length-of-stays greater or less than $t = 2$ days,

[4] `https://github.com/YerevaNN/mimic3-benchmarks`

Example data:{'static': shape: (1, 18)

gender_0	gender_1	marital_status_0	marital_status_1	…	race_1	race_2	race_3	race_4
--- i16	--- i16	--- i16	--- i16		--- i16	--- i16	--- i16	--- i16
0	1	1	0	…	0	0	0	0

, 'dynamic_0': shape:

Anion Gap	Bicarbonate	Calcium, Total	Chloride	…	Red Blood Cells	Sodium	Urea Nitrogen	White Blood Cells
--- f64	--- f64	--- f64	--- f64		--- f64	--- f64	--- f64	--- f64
0.31	0.47	0.36	0.5	…	0.41	0.88	0.09	0.06
0.31	0.47	0.36	0.5	…	0.41	0.88	0.09	0.06
0.31	0.47	0.36	0.5	…	0.41	0.88	0.09	0.06
0.31	0.47	0.36	0.5	…	0.41	0.88	0.09	0.06
0.31	0.47	0.36	0.5	…	0.41	0.88	0.09	0.06
0.31	0.47	0.36	0.5	…	0.41	0.88	0.09	0.06
0.29	0.41	0.39	0.58	…	-1.0	0.91	0.11	-1.0
-1.0	-1.0	-1.0	-1.0	…	0.39	-1.0	-1.0	0.06
-1.0	-1.0	-1.0	-1.0	…	0.39	-1.0	-1.0	0.06
-1.0	-1.0	-1.0	-1.0	…	0.39	-1.0	-1.0	0.06

,

Diastolic blood pressure	Heart rate	Oxygen saturation	Respiratory rate	Systolic blood pressure	Temperature
--- f64	--- f64	--- f64	--- f64	--- f64	--- f64
0.41	0.55	-1.0	0.1	0.58	-1.0
0.41	0.55	-1.0	0.1	0.61	-1.0
0.41	0.55	-1.0	0.1	0.61	-1.0
0.41	0.55	-1.0	0.1	0.61	-1.0
0.39	0.51	-1.0	0.11	0.56	0.56

}

Figure 12: Example stay with static feature table and time-series data table(s) from ED and the first $t_{max} = 48$ hours during the HOSP stay, after cleaning and processing. Note: some feature columns are hidden.

since 2 days is a commonly used threshold to determine severity throughout the NHS. We applied one-hot encoding to categorical data such as gender, insurance type, marital status and race. In particular, we group the MIMIC-IV race feature values into 5 broader categories: White, Black, Hispanic, Asian, Other/Unknown. For continuous variables such as age, we used min-max scaling to standardise the values to the range [0,1]. For time-series data to support parsing through to models such as the long-term short memory network (LSTM), we applied resampling to turn the infrequent data into regular intervals. This was achieved by upsampling to 1-minute intervals and forward filling, before downsampling to a desired frequency f with a window-wise average. Any remaining missing values were imputed with a second forward filling step.

3.2.3 Handling multiple event tables

Time-series data was obtained from both lab events and vital signs, corresponding to data from the hospital stay and emergency department respectively. Due to this we used different sampling frequencies depending on the average time between readings across each stay. As described in the MIMIC documentation, vital signs data are routinely collected every 1-4 hours, so we set $f_{ED} = 30$ minutes. On the other hand, lab events are typically recorded 1-3 times daily, therefore we set $f_{HOSP} = 5$ hours.

3.3 Multimodal Modelling

We mirrored the approach of [57] by creating a fusion model in which the time-variant and invariant data are fused with a simple fusion approach. In this case, we parse each modality through an independent framework, before combining their representations in a final fusion layer. We treat the stay-level data as being comprised of three distinct subsets: time-invariant data (static), and two time-varying series (vital signs and lab events). Whilst the time-series data could be concatenated and treated as a single series, we chose to keep them separate, due to their differing frequencies and to avoid introducing more missingness whereby lab events are not recorded at the same time as vital signs. In some cases the gap between the last vital sign and first lab event could be so large that this will lead to the introduction of several hours with no associated data. A schematic summarising the final model is shown in Figure 13.

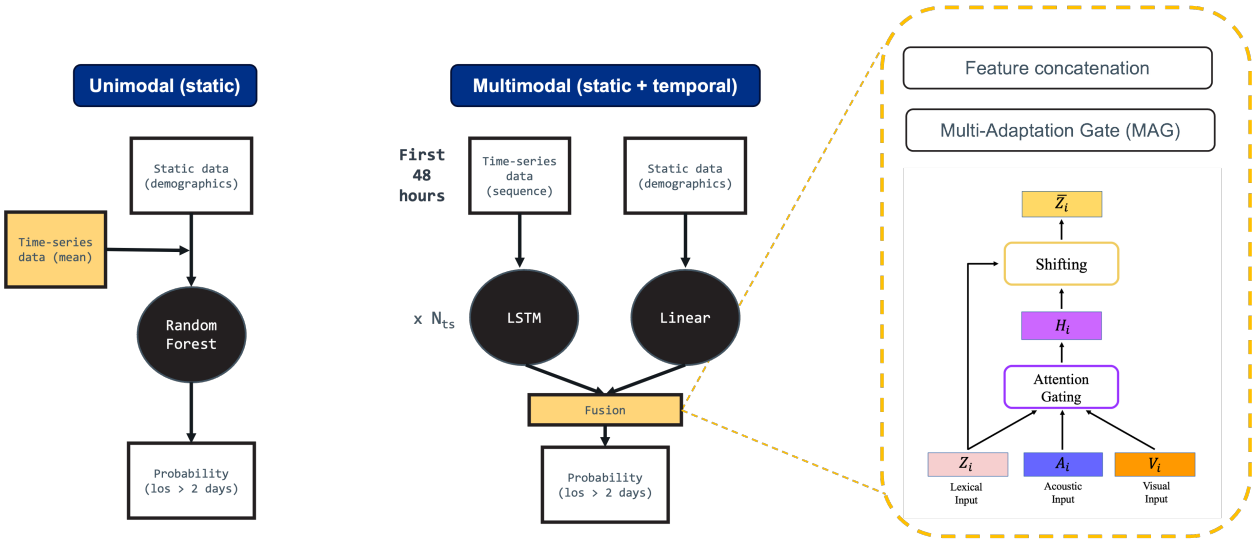


Figure 13: Overview of the models: unimodal, multimodal with two fusion methods: concatenation and a multi-adaptation gate (MAG).

3.3.1 Time-invariant feature representations

Different approaches can be taken for modelling the time-invariant (static) features, which include age, gender, ethnicity, martial status, los (in ED) insurance type. Here, we adapt the approach of Yang and colleagues [57] and use a fully-connected layer to embed these features. However, it is worth noting alternative approaches such as [56], which used a Transformer to embed each feature independently before parsing to the fusion layer. This has the advantage of learning separate embeddings for each feature, and removes the requirement to apply one-hot encoding. However, in this project we are more interested in modality-level explainability i.e., when certain modalities become more or less useful to prediction (time-series > clinical notes). Therefore we have decided to follow the approach of Yang to avoid introducing more complexity into the model and an additional explanation step.

3.3.2 Time-varying feature representations

For the time-varying data, we use an Long Short-Term Memory network (LSTM) [24] to generate representations prior to fusion. LSTMs are a type of recurrent neural networks (RNNs) that excel at capturing long-range dependencies in time-series data. LSTMs have a unique architecture that includes a cell state and multiple gated mechanisms: the forget gate, input gate, and output gate. These gates retain, update, and output information at each time step, enabling the network to leverage relevant historical data. Alternative architectures include Transformers [53], which unlike LSTMs can leverage long-range relationships within data. Examples have applied Transformers to electronic health records [30, 50, 58]. However, as an initial approach, we chose to first utilise LSTMs as they are typically less complex with less trainable parameters. We trained two separate LSTMs for the time-series sources, due to their inherent differences in frequency and feature variability.

3.3.3 Dealing with Missingness

Missingness is evident in two ways: time-wise and feature-wise. We use feature-wise missingness, to refer to entries in which not all types of reading are measured at a given time. For example, respiratory rate may be recorded, but not blood pressure. Moreover, time-wise missingness is used to refer to the irregular gaps between readings, which varies between patients.

Structured missingness refers to a cases where data is not missing at random but rather in a systematic way, often due to underlying factors or decisions made during the data collection process. This can lead to missing data being correlated with other observed variables. As such, understanding structured missingness can play a role in identifying bias where certain groups of patient may have more missingness than others. This can also cause the time-series (sequence) length to vary between patients. This could introduce unwanted bias where the model learns better from patients with longer series (more data, more information).

To address feature-wise missingness, we rely on research from Lipton and colleagues who highlight that using a meaningless value can be effective way of encouraging models to learn when/how to ignore certain features [33]. Alternatively strategies include imputation e.g., with the mean value or introducing an additional "missingness indicator" flag, to mark rows with missingness [31, 19]. Whilst the codebase supports several imputation strategies, here we work on a simple imputation strategy where we use an 'meaningless' value (-1) to fill feature-wise missingness as discussed earlier. For time-wise missingness, we have made use of torch's `pack_padded_sequence` function to allow parsing batches of different time-series lengths to the LSTM embedding layer. This prevents the need to truncate or add additional empty rows to ensure all series are of a fixed length.

Despite this, within a given time-series some stays will have more or less time-wise missingness as a result of resampling. We investigated this further during the bias and fairness exploration stage.

3.3.4 Multimodal Fusion

The learned representations of the static and dynamic features can be fused in a variety of different ways (see Section 2.1.1). The `fusilli` [52] package provides an out-of-the-box solution to compare multiple fusion strategies against one another. However, as of writing, this software is limited to tabular-tabular or image-tabular fusion, and is not trivially adaptable for time-series data. Therefore, as an initial approach we decided to compare two fusion methods in this project: concatenation and multi-adaptaion gate fusion.

Concatenation is a naive baseline approach which simply combines the representation vectors from each modality into a single vector which can be turned to a probability output node via a learnable mapping. Multi-adaptation gate fusion was proposed by Rahman and colleagues [43] as an attachment for the BERT models, to control how much information should be taken from each modality during fusion. This operation is performed at the word level to shifting the word representation according to contextual information from the other data sources. A schematic diagram can be seen in Figure 14. Extensions have been developed such as MAG+, which extend the 1-layer network to a dynamic, cross-attention based approach [60]. More specifically, Yang and colleagues [57] provide an adapted version of MAG for EHR data where the modalities are inherently asynchronous, unlike in the original study. Due to this fusion is performed on the sample level instead of word or feature-wise.

Here, we use MAG to fuse representations from three modalities to compare against the concatenation method. This is motivated by the uniqueness of the MAG to allow for a certain data type to be chosen as the primary modality, Z, with all others as accompanying ones providing additional context. This is achieved by shifting Z by the a displacement vector H to obtain the new fused version, $\bar{Z}$:

$$H = g^a(W_a \cdot A) + g^v(W_v \cdot V) + \beta_H \tag{1}$$

$$\bar{Z} = Z + \alpha H \tag{2}$$

A and V correspond to acoustic and visual inputs used in the original paper (see Figure 14) however in our case are the two time-series representation vectors. The learned gating vectors g, provide a mechanism for highlighting the relevant information in the two accompanying modalities whilst being conditioned on the primary. The parameter α can be used to control how much shifting is applied to the primary modality and β is a randomly bias term.

We implement MAG by adapting the code used by Yang and colleagues [57]. In this implementation the gating vectors, g, are defined by a Linear layer:

$$g_i = \texttt{RELU}(\texttt{Linear}(Z, Y_i)) \tag{3}$$

where Y_i is any auxiliary modality and

$$H = \texttt{Linear}(g_1 Y_1, g_2 Y_2) \tag{4}$$

such that the main modality Z can be displaced according to Equation 2. This is followed by normalisation and dropout operations. We also add flexibility to support either one or two auxiliary modalities.

3.3.5 Clinical notes representation

As an extension, we also consider the incorporation of text data by leveraging MIMIC-IV discharge summary notes. To avoid inclusion of information indicative or related to the length-of-stay, we focus on the section related to 'History of Present Illness'. A snippet from an example patient discharge note is shown in Figure 15.

Each snippet was tokenised and embedded using a publicly available, pretrained BERT model`Bio+Discharge Summary BERT` [7] available via the `spacy` package. This model was trained specifically on ICU clinical discharge

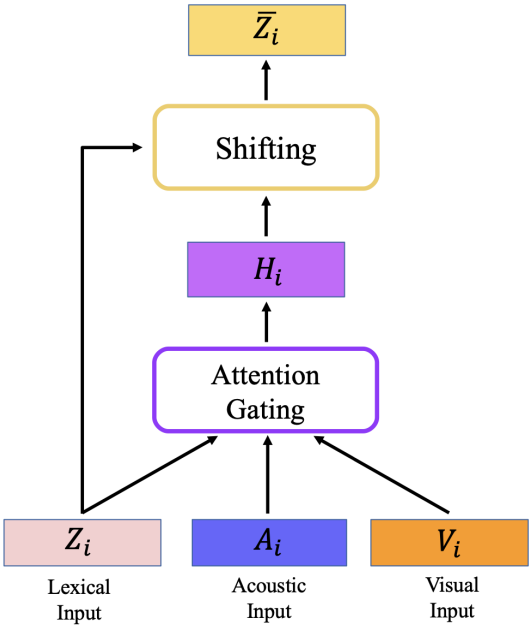


Figure 14: The multi-adaptation gate mechanism was proposed by [43] and the figure has been taken from the paper.

summary notes from MIMIC-III. As a proof of concept, we used a max length of 128. To convert each snippet to a single embedding, we averaged over the embeddings for all sentences.

3.4 Explainability, bias and fairness

This project seeks to investigate explainability, bias and fairness in the context of multimodal modelling for healthcare. As such, whilst is important to obtain a useful, accurate underlying model in order to generate meaningful explanations and insights, our the goal was not to focus on performance optimisation. Therefore, once we obtained a set of models with 'good' accuracy on our predictive task, we sought out to apply explanation methods to understand model behaviour.

To understand which features were most informative, we applied `captum`'s[5] implementation of SHapley Additive exPlanations (SHAP) . SHAP works by attributing each feature's contribution to the final output and is based on the concept of Shapley values from cooperative game theory [34]. SHAP assigns a value to each feature that reflects its importance by considering all possible combinations of features. This ensures a fair distribution of the predicted value among the features, capturing the effect of each feature both individually and in interaction with others. One of the key strengths of SHAP is its model-agnostic nature, meaning it can be applied to any machine learning model (see Section 2.2). These highlight which of the static features are most influential using the learned weights of both the Random Forest and LSTM networks. This provides some level of explainability as we can compare the order (and magnitude) of the attributed importance across the different models to begin to unpick the models' behaviour.

Aside from feature-level importance, we were also interested in understanding how the modalities are combined. The weights assigned to the fusion layer can be used to shed light on how the relative importance time-series and static data. For the concatenation approach, this is straightforward as the weights from the fusion layer directly correspond to the each input modality. However, for the MAG module, the meaning of the learned weights is less transparent. To unpick the influence of each modality for these fusion models, we instead compare the relative order of features depending on the chosen primary modality.

Being able to explain the model forms one piece of the puzzle for moving towards fairness in AI. This is due, in part, to the ability for XAI to shed light on biases that may be present in the model. By comparing the explanations for the different multimodal and unimodal models we also also investigated the presence of model bias which can in turn, influence how the features are combined to generate a final prediction. For example, we hypothesise that a model prone to bias e.g., if missingness linked to a particular subset, then the relative importance of time-varying features

[5] `https://captum.ai`

History of Present Illness:
Patient is a ___ with history of coronary artery disease c/b ischemic MR ___ DES to LCX ___, TTE ___ with mild regional LV systolic dysfunction), heart failure with preserved ejection fraction (LVEF 50% ___, peripheral vascular disease, chronic kidney disease (stage IV), prior unprovoked DVT c/b severe UGIB while on AC, HTN, dyslipidemia, and T2DM who presents with several days of shortness of breath.

Patients says that she first noticed rather acute onset dyspnea starting ___ when trying to walk up the stairs in her home.

She had to sit down and catch her breath, whereas just days prior
she was able to mount ___ of stairs without difficulty.
Patient denies any associated chest pain or palpitations. No dizziness or lightheadedness. Patient further denies any cough, fevers/chills, or pleuritic chest discomfort. She has not experienced any symptoms consistent with orthopnea or PND. No increased ___ swelling, patient notes that she has experienced this in the past.

Patient takes her weight nearly every day, 7lbs reported weight gain over the past week (154lbs -> 161lbs), which she attributes to eating more over the ___. She is currently
taking torsemide 40mg qd, no missed doses. No issues with abdominal bloating or constipation. No recent travel.
Patient's
husband just recovered from a viral URI.

Figure 15: Snippet from an example discharge summary note for a patient in the MIMIC-IV dataset.

may be less than static ones for subjects in that subset. We also evaluated our models using fairness metrics available via the Fairlearn[6] package. Hardt and colleagues [21] proposed equalised odds as a better alternative to demographic parity particularly that ensures fairness across all subsets. We report both values (and their ratios) in our results for completeness, however the limitations of each fairness metric should always be considered when interpreting their values.

3.5 Technical implementation

The codebase[7] was designed to support different input tables and processing requirements. Therefore, it currently supports the following imputation strategies: mean, mask, value, None. Additionally, when processing the data, there is also a choice to stratify the sample according to the length-of-stay threshold. This ensures that the number of positive and negative cases are roughly balanced. Moreover, the specific event items extract can be specified via a text file, as opposed to the default top 20 lab items. Feature scaling and resampling can optionally be turned off, and the maximum time t_{max} can be changed from the default value of 48 hours after hospital admission.

Most significantly, the `polars` library has been leveraged to facilitate efficient reading and writing from the data. Due to the large number of patients and events present in the data ($\sim$100K and $\sim$110M respectively), a key challenge lay in being able to read from the downloaded csv's. Using `polars` allowed the time taken to extract and preprocess all desired stays to be reduced from several days (compared to using `pandas` only), to approximately 5 minutes.

Model training and evaluation was performed on a single Microsoft Azure machine running a Linux Server with the following specifications: 1 x NVIDIA Tesla T4 GPU and 4 x vCPUs (28 GiB memory).

4 Results & Discussion

4.1 Data Analysis

Before training any models, we sought out to characterise the data with some initial analysis, shown in Table 2. This highlights that there is a slight imbalance in the age distribution of those with longer stays, who tended to be older.

[6] `https://fairlearn.org`
[7] `https://github.com/nhsengland/mm-healthfair`

Whilst this is intuitive, it helped to inform some further analysis into whether age is correlated with other features. We also visualised the Figure 16 and correlation between the different demographic features in Figure 17.

Feature	LOS < 2	LOS > 2
Age (mean)	55.8	62.3
ED (mean - hrs)	0.37	0.32
Gender (frac. f)	0.75	0.75
# of stays	7660	7687

Table 2: Characteristics of the (training) dataset.

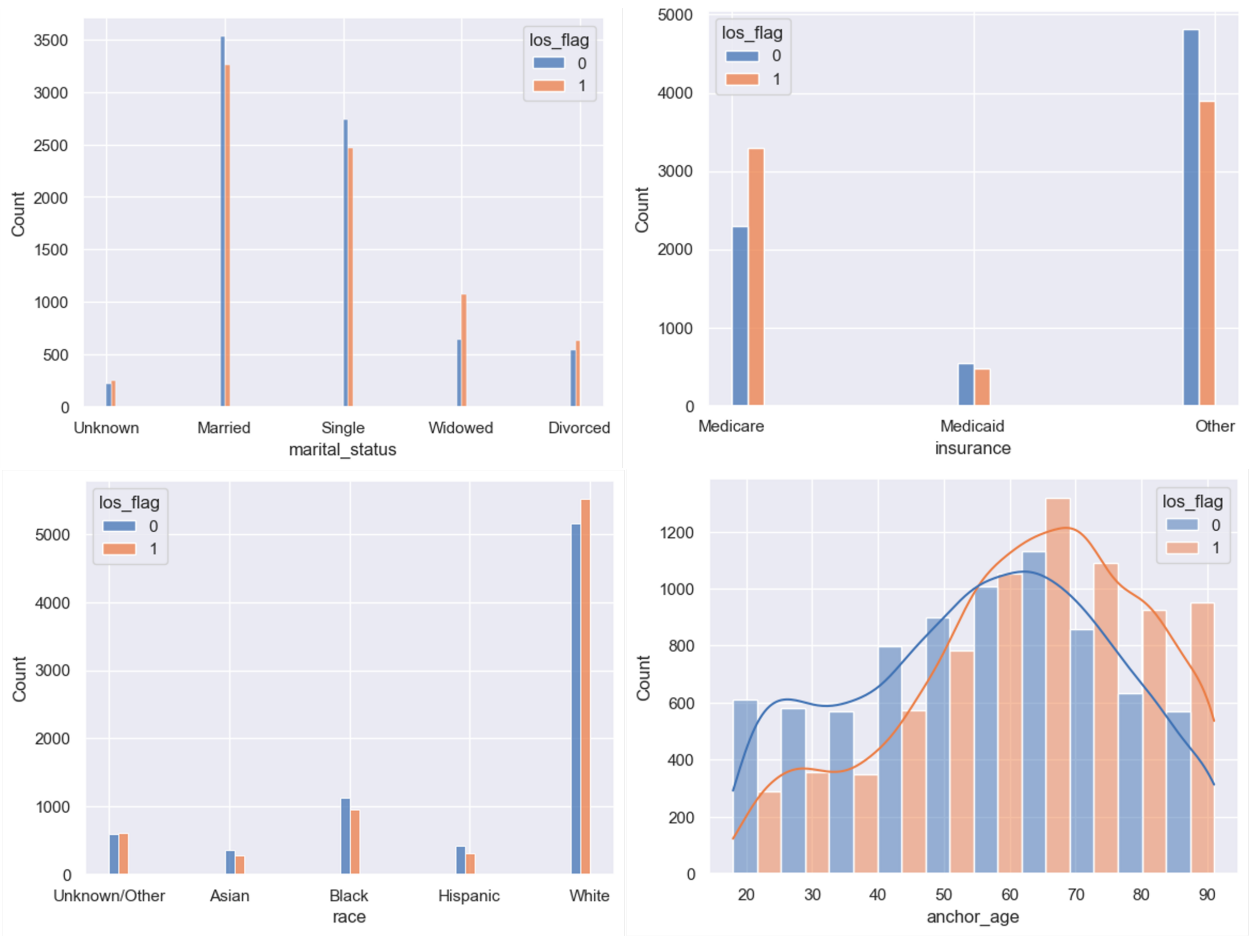


Figure 16: Histograms of categorical features from the (training) dataset stratified by LOS.

These visualisations highlight the presence of imbalance in the data, according to the different race groups, marital status and insurance type. For example, the majority of stays are from White patients. There are also few subjects with Medicaid insurance as opposed to Medicare, although it must be noted that the majority patients were 'Other' insurance types besides these. We chose to focus on this aspect, since insurance type can be a influencing factor in the interaction of the public with the Healthcare Service in the data. The correlation plot also highlights the relationship of the LOS indicator with age and insurance type most notably (-0.13 and 0.17) respectively. This supports the notion that there is a higher likelihood of a person having a longer stay on Medicare than Medicaid, which could be a labelling or sampling bias.

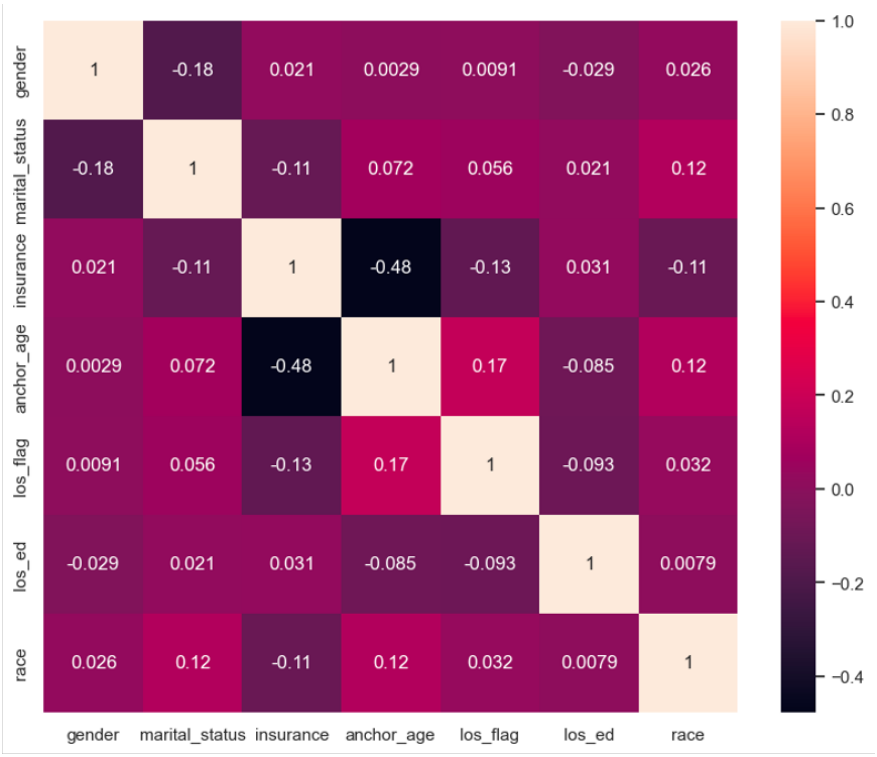


Figure 17: Correlation heatmap for categorical features in the (training) data.

4.1.1 Quantifying missingness

To better understand the amount of missingness, we also computed the amount of missing values for each feature stratified by LOS. Figure 18a) shows that certain features such as Glucose, Anion Gap, Magnesium and Platelet Count, have the most missingness. There is also a bias towards LOS, where some features have more missingness depending on the LOS. However, as we decided to follow the strategy as in [33], the model should learn to ignore these meaninglesss values during training.

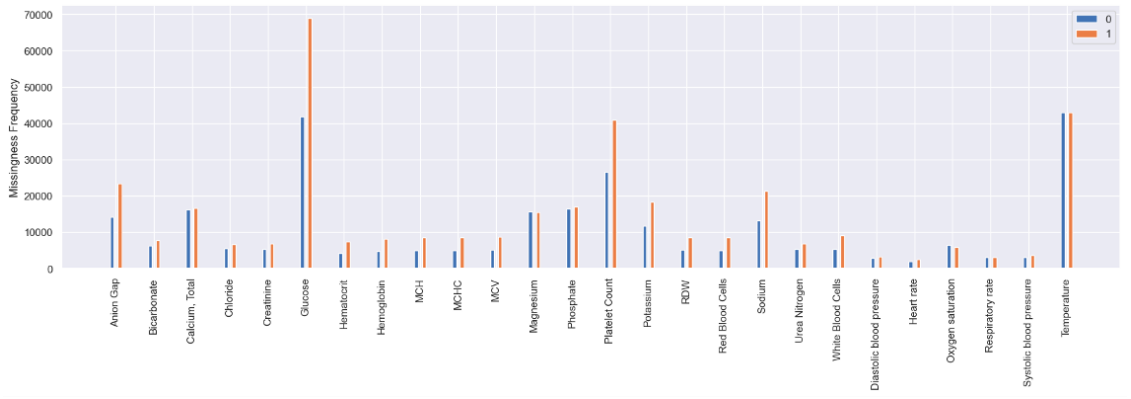
On the other hand, the time-wise missingness introduced as a result of resampling can also exhibit a link to protected characteristics. This mimics the idea that certain patients have more or less 'information', and therefore more or less recorded measurements. In a real life setting, this could arise due to social or human biases which cause certain groups to spend less time in hospital, or receive differing levels of attention during the course of their stay. To quantify this, we computed the percentage of missingness by calculating the fraction of non-informative timepoints in the series:

$$f_{\text{missing}} = \frac{L_{\text{upsampled sequence}} - L_{\text{initial sequence}}}{L_{\text{upsampled sequence}}}$$

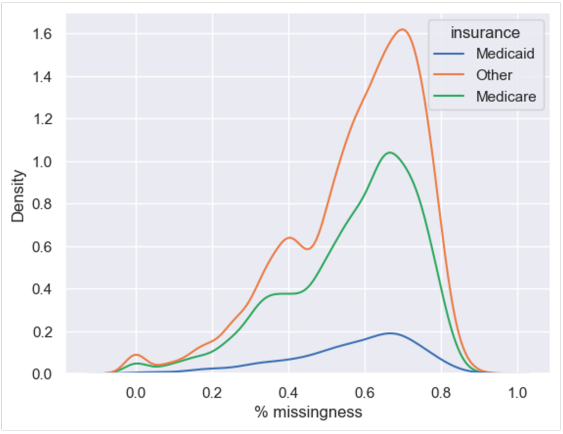
Figure 18b) and c) reflect that whilst there is a slight increase in the amount of time-wise missingness for the Other and Medicare groups, this is most substantial when stratifying by LOS.

4.2 Assessing the importance of time-dependant information

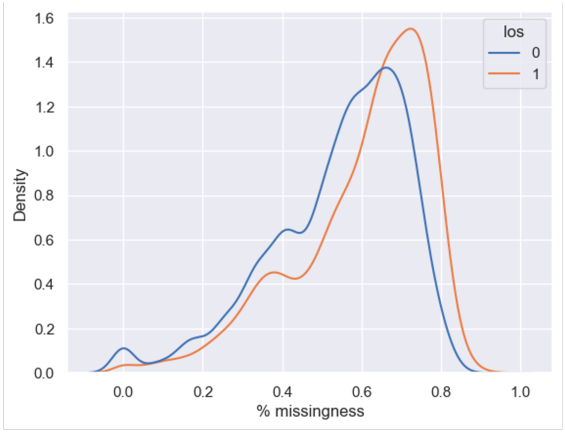
First, we set out to compare the performance of unimodal models with multimodal networks. This involved comparing a model trained only on static data to the different fusion models to understand whether temporal data contains useful information for prediction. We used a Random Forest to train two unimodal models based on static data only. One model (static only) was trained on demographic information only, whereas the other (static + mean) included the average value for the various dynamic features from lab events and vital signs. In contrast, three fusion models were trained and evaluated using concatenation and MAG. For the MAG approach, we also included a model where the primary modality was the time-series data (lab events) to see if this had an impact. The performance results for predicting LOS greater than 2 days (positive) as shown in Table 3.



(a)



(b)



(c)

Figure 18: Missingness in the (training) data for dynamic variables. (a) Frequency of missing values across features stratified by LOS. Percentage of time-wise missing values introduced during resampling stratified by (b) insurance type and (c) LOS.

Model	ACC	BACC	AUROC	AUPRC
Baseline	0.587	0.587	0.619	0.582
Dynamic	0.693	0.692	0.772	0.780
Fusion (Concat)	0.833	0.833	0.899	0.899
Fusion (MAG)	0.831	0.830	0.893	0.892
Fusion (MAG-TS)	0.818	0.818	0.890	0.890

Table 3: Performance results for the five different prediction models.

These results indicate that the information present in vital signs and lab events contains useful information for prediction, leading to a $\sim$10% improvement in predictive accuracy when the mean values were added alone. More importantly however, all three fusion models outperform the unimodal approaches, highlighting the gain in useful information when time-varying features are included.

Surprisingly, concatenation led to the best accuracy overall, though the differences between all three approaches were not substantial. Despite this, upon inspecting the training curves for each fusion model shown in Figure 19, we identified that the model trained with the time-series as the primary modality converged much quicker than the other models. This may be indicative of the ability for the model to identify meaningful patterns from the time-varying features more quickly.

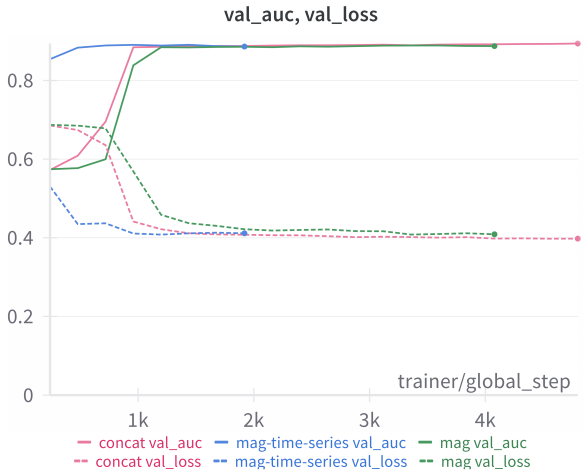
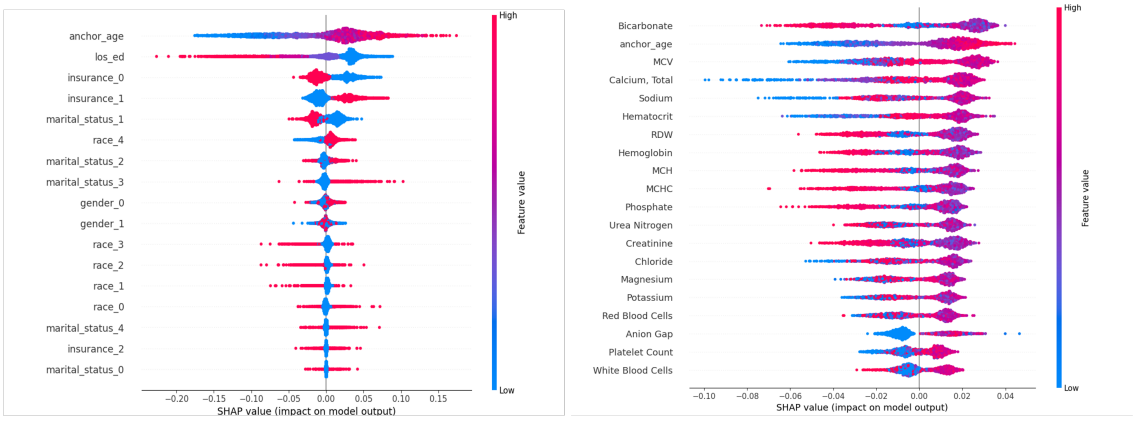


Figure 19: Validation loss and accuracy curves per epoch for the three different fusion models.

4.3 How do multimodal modelling decisions impact on explainability?

The focus of this work was not to optimise performance, but rather unpick the interplay between multimodal modelling approaches, fairness and explainability. Figures 20a) and b) provide an insight into how how the features are driving the prediction for the unimodal Random Forest models, averaged over the test samples. In both cases only static features were included, however comparing them highlights the relative importance of the different data sources. For the baseline model, age and LOS in the emergency department (los_ed) have the largest influence on the model output. Specifically, older ages and those with a shorter stay in the ED are associated with longer hospital stays. This supports the findings during the data analysis stage, where los_ed and age were amongst the top 3 features correlated with LOS in hospital (see Figure 17). Introducing the mean value for the dynamic features across lab events and vital signs, led to an increase in performance and also impacts the order of feature importance. The summary beeswarm plot highlights that predictions are driven by the bicarbonate (C02) blood levels, followed by age and Mean Corpuscular Volume (MCV), a measure of red blood cell size. This suggests that poorer blood health, is indicative of longer stays in hospital. However more generally, the overall influence of all features is more balanced in when including dynamic features, which indicates that there is more useful information in these features.



(a) (b)

Figure 20: SHAP beeswarm plots for the unimodal Random Forest models. a) baseline (demographic only) (b) dynamic (demographic + time-wise features mean) model.

Moreover, for the fusion models, we can also apply SHAP using the learned submodels for each input modality. Here, we focus on the LSTMs trained independently on time-series from lab events and vital signs data. Figure 21 reveals that there is little difference between the three fusion models in the driving features, particularly for the lab events

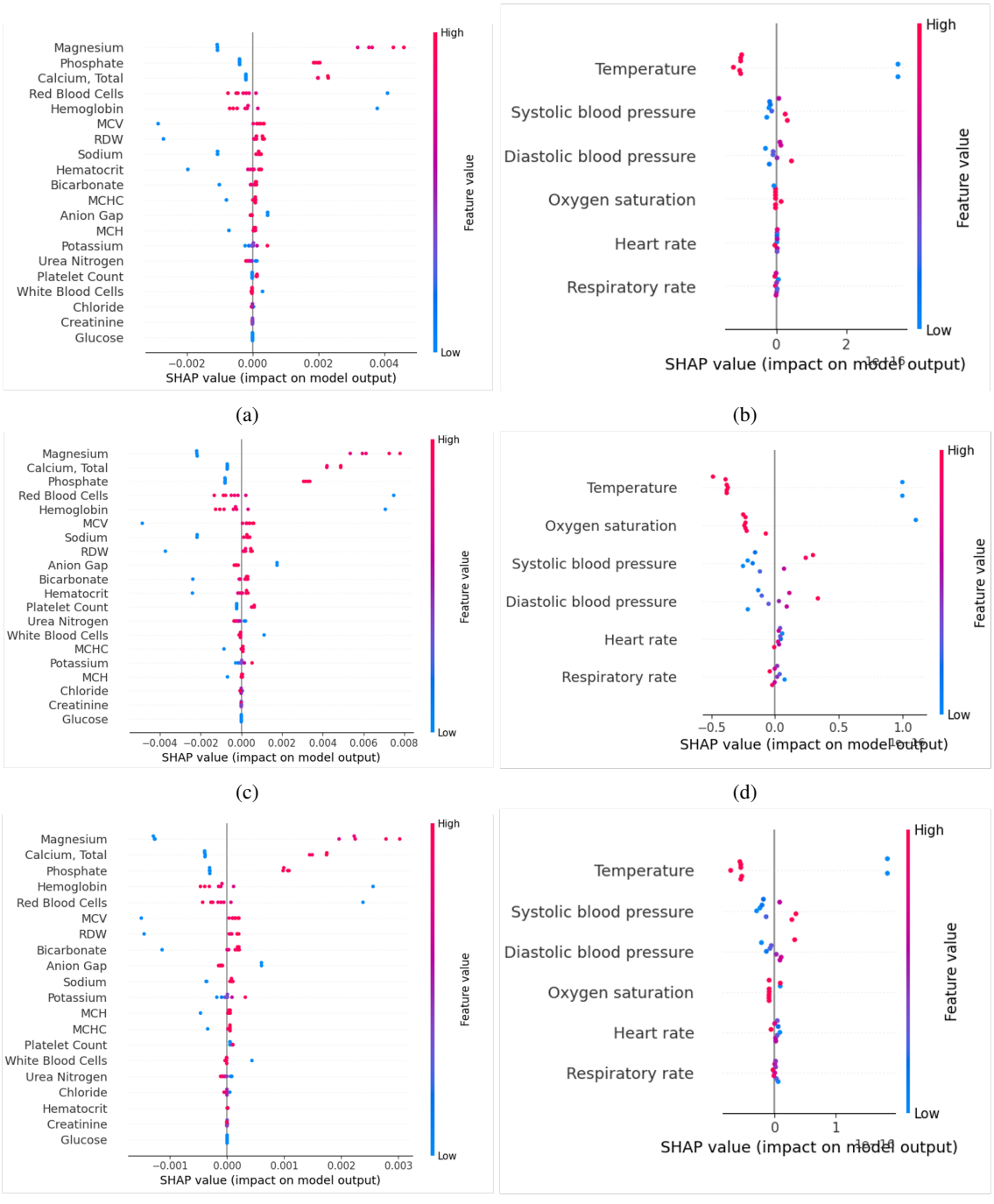


Figure 21: SHAP beeswarm plots for the fusion models. Panels (a), (c) and (e) show the influence of lab events features and (b), (d), (f) show the influence of vital signs for the concatenation, MAG and MAG-TS models respectively.

features, despite there being differences in absolute performance. This consistency is reassuring, as it suggests that feature importance is independent of fusion choices. For vital signs feature values, the order of the top four driving features does vary across the different models. Most interestingly, when time-series is the primary modality (Figure 21f) the blood pressure values become more indicative of LOS compared to the MAG fusion model with demographic information as primary (Figure 21d). There is also a clear difference in the SHAP values attributed to oxygen saturation levels between these models, where it is more discriminative for the MAG model compared to the MAG-TS model,

which agrees more closely with the concatenation approach. In this way, comparing the explanations across different models is useful for understanding the stability of model explanations, but can also reveal insights into the impact of fusion choices.

4.4 How do modelling decisions impact on fairness?

We computed the equalised odds (and difference) and demographic parity ratios for all models across four protected characteristics: gender, race, insurance type and marital status. These values are provided in Table 4. We report the minimum ratio (difference) between all subgroup pairs. However, interpreting these results is non-trivial, since most groups features have more than 2 unique values (with the exception of gender). We also visualise the difference in Figure 22. This displays the model accuracy and false negative rates for different subgroups based on protected characteristics across each model.

These results reflect the quantitative findings of Table 4, that the model is most unfair with respect to insurance type and most fair for each genders. This could be due to the presence of bias in the training data itself, as seen in Figure 4.1. In contrast, despite the heavy imbalance with respect to race, the difference in accuracy and fairness metrics were still better than with respect to insurance type. This suggests that it is not only about bias present in the data, but also the underlying relationship between the label and feature.

Focusing on insurance type alone, we can also unpick how modelling choices affect model fairness. Quantitative results highlight that the fusion models produced fairer models with respect to insurance type. Whilst accuracy and false negative rates were consistent across the three fusion models in this case, the concatenation method provided the lowest EOD based on insurance type. This could be due to the lower number of parameters present in the concatenation model, which may make it less prone to overfitting and therefore bias towards imbalanced samples.

Model	**EOR ↑**				**EOD ↓**				**DPR ↑**			
	Gen.	Race	Insur.	Mar.	Gen.	Race	Insur.	Mar.	Gen.	Race	Insur.	Mar.
Baseline	0.98	0.47	0.25	0.36	0.01	0.26	0.63	0.56	0.99	0.52	0.32	0.44
Dynamic	0.87	0.67	0.45	0.44	0.04	0.11	0.24	0.25	0.94	0.70	0.56	0.62
Fusion (C)	0.83	0.73	0.87	0.67	0.05	0.05	0.06	0.06	0.95	0.82	0.75	0.75
Fusion (M)	0.82	0.64	0.84	0.50	0.06	0.06	0.07	0.11	0.94	0.77	0.73	0.72
Fusion (M-TS)	0.93	0.77	0.60	0.53	0.06	0.04	0.09	0.12	0.93	0.83	0.72	0.73

Table 4: Fairness metrics evaluated for the five models. EOR = equalised odds ratio. EOD = equalised odds difference. DPR = demographic parity ratio. C = Concat. M = MAG. M-TS = MAG-TS. Gen. = Gender. Insur. = Insurance. Mar. = Marital status.

4.5 Exploratory work

4.5.1 Incorporating clinical notes

We also began to investigate how to incorporate text as an additional modality in this framework. The clinical notes available as part of the MIMIC-IV dataset provide summaries of a patients medical history. Whilst we were able to generate embeddings for all stays, we ran into computational limitations as the amount of required storage space drastically increased. Loading the notes and tabular data for each stay during training led to out-of-memory issues. We circumvented this by generating the embeddings before training (during preprocessing) and loading the embedding vectors during training directly, however this still led to a significant increase in training time. Moreover, generating a single embedding for an entire paragraph may be too coarse to provide meaningful information, in line with the dimensionality of other features.

4.5.2 How does the presence of model bias impact explanations?

Given that the model we trained displays some evidence of bias towards insurance type, we designed a simple ablation study which aims to generate a more biased model using subsampled data. We defined 'high' and 'low' missingness based on the fraction of time-wise missingness described in Section 4.1.1, where high missingness as greater or less than 40%. Figure 23 highlights that whilst all groups have more stays with 'high' missingness, the skew is strongest for those with Medicare or Other insurance. Using this, we generated a sampled dataset by exaggerating the skew such that patients on Medicaid had more missingness compared to patients on Medicare, and (Other category remained the same as the original). This meant filtering stays such that the proportion of missingness for Medicaid patients was higher

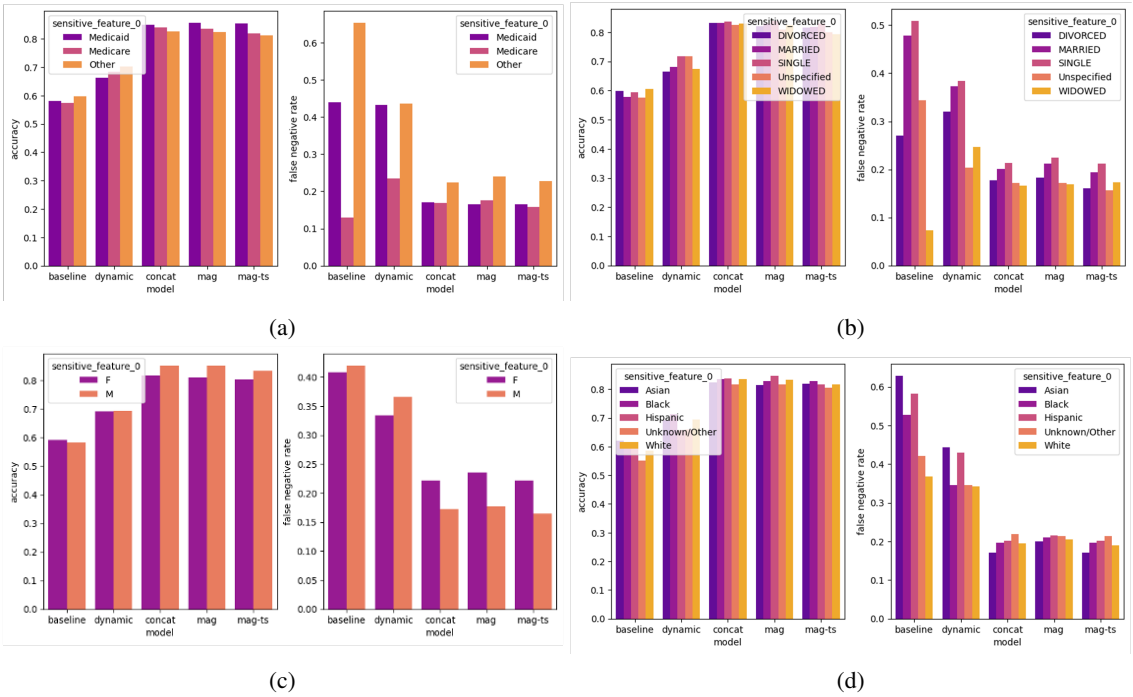


Figure 22: Accuracy and false negative rate shown across each subgroup for each protective characteristics across the five different models.

(∼70%) than those on Medicare (∼30%). For a fair comparison, we use the original distribution to create dataset with the same sample size whilst filtering stays at random. In both cases, we ensured that the balance between long and short stays (the target variable) remained the same.

By observing differences in fairness and explainability outputs for models trained on these datasets with different levels of bias, we can begin to understand how data bias manifests in this context. For example, does a model with more bias due to structured missingness as described, rely more heavily on static features? However, due to the low numbers of patients on Medicaid overall, the final sampled datasets were limited and insufficient for training a model effectively.

4.5.3 Bias mitigation

Fairness metrics can be used to derive insights into the presence of bias, but how can we prevent or avoid this? Simply quantifying bias is not enough for ensuring safe and ethical deployment of AI models en masse. Bias mitigation strategies have been proposed as a way to reduce or prevent model bias as described in Section 2.3. We began to explore whether we could debias an already biased model based on our earlier findings. The Fairlearn package supports post-training strategies such as adversarial learning and threshold optimization. Due to time constraints, our analysis were limited to threshold optimization, a technique which adjusts the model decision boundary to optimise based on a chosen fairness metric. For a proof-of-principle, we applied to this to one of the existing models, which from our fairness analysis shows indications of bias towards insurance type. Figure 24 shows the optimal threshold to balance balanced accuracy score across the three subgroups of insurance type best, achieving an overall false negative rate of approximately 0.2. This is a crude and simple approach to model debiasing but it is an active and rapidly evolving area of research. For example, Nam and colleagues propose a strategy to debias models by learning from failure, whereby they hypothesise that biases are most likely to be leveraged for the more 'difficult' to classify cases. Using this, they exploit these failure cases in order to understand and reduce bias.

For increased clinical applicability, it may be more useful to debias networks during training. For example, besides adopting fairness metrics post-hoc, an alternative approach would be to incorporate fairness constraints during training, as shown by Kim and colleagues who propose fairness-aware multimodal learning for automatic video interview assessment [28]. Here, a fairness loss is used in combination with an accuracy loss (adversarial training) to ensure Strong Pairwise Demographic Disparity (SPDD). For instance, Bahng and colleagues show the carefully adapted loss functions can be used to mitigate against such bias during training [9].

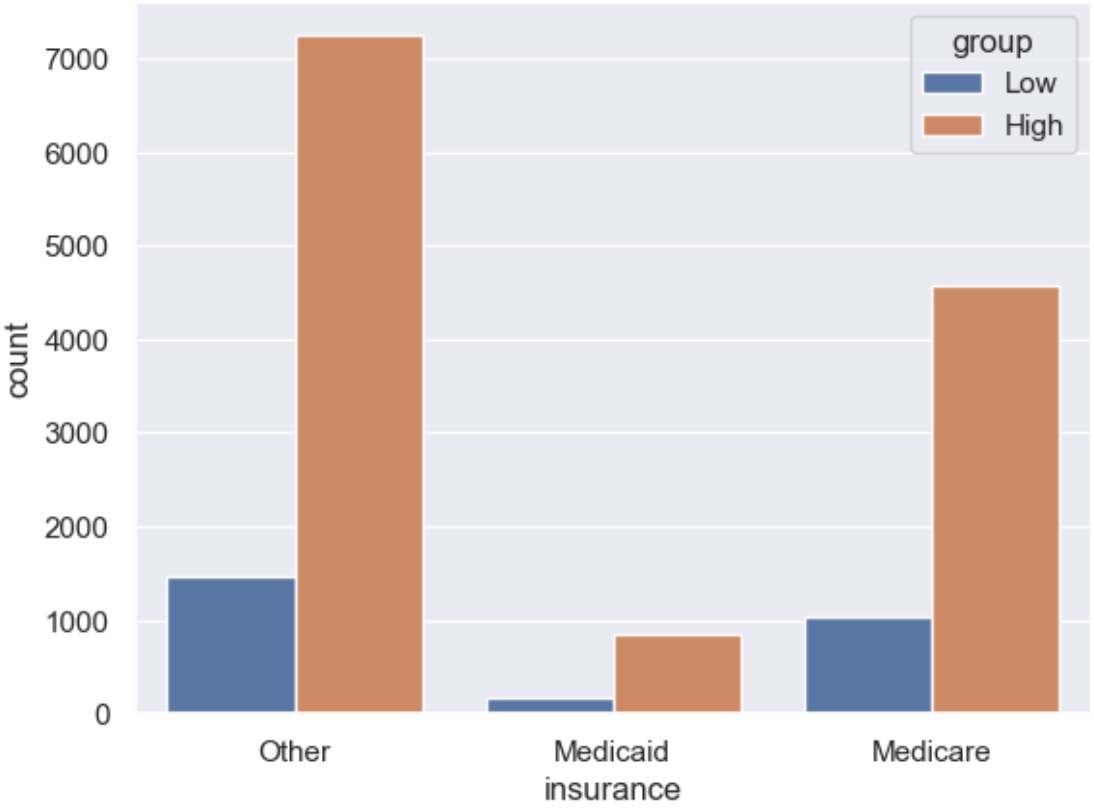


Figure 23: Number of stays across different insurance types split based on high and low missingness.

5 Future Work

This work demonstrates the utility of different fusion strategies for length-of-stay prediction, leveraging time-series data specifically. We present initial findings which combine explainability and fairness metrics for the exploration of model bias. In our exploratory analysis, we describe a set of experiments which aim to tease out further insights into how an increase in data bias manifests in model explanations. To continue building on these results we provide a few opportunities that future work could focus on:

5.1 Synthetic data generation

We were limited by the sample size, particularly when exploring bias with respect to insurance type. To generate quantifiable insights into how increases in data or model bias manifest in downstream outputs, it would be beneficial to have greater control over the dataset attributes. This could be achieved with synthetic data generation, in order to increase the size of certain subpopulations and systematically compare models with and without bias. Some efforts have been made towards this, with the release of open-source data generation models such as Synthea[8][54].

5.2 Transformers for sequence modelling

Here, we decided to use LSTMs to generate vector representations of dynamic vital signs and lab events. However, existing studies have demonstrated that Transformers are able to, in some cases, outperform LSTMs by capitalising on long-range patterns in temporal data. Although the focus of this project is not to achieve optimal performance, future work could seek to investigate whether Transformers provide a significant advantage over LSTMs. Moreover, some of these models are open-source, and pretrained models could be leveraged in future work to boost predictive power. For example, the MOTOR model [50] is a time-to-event foundation model trained on healthcare data which can be

[8] https://syntheticahealth.github.io/synthea/

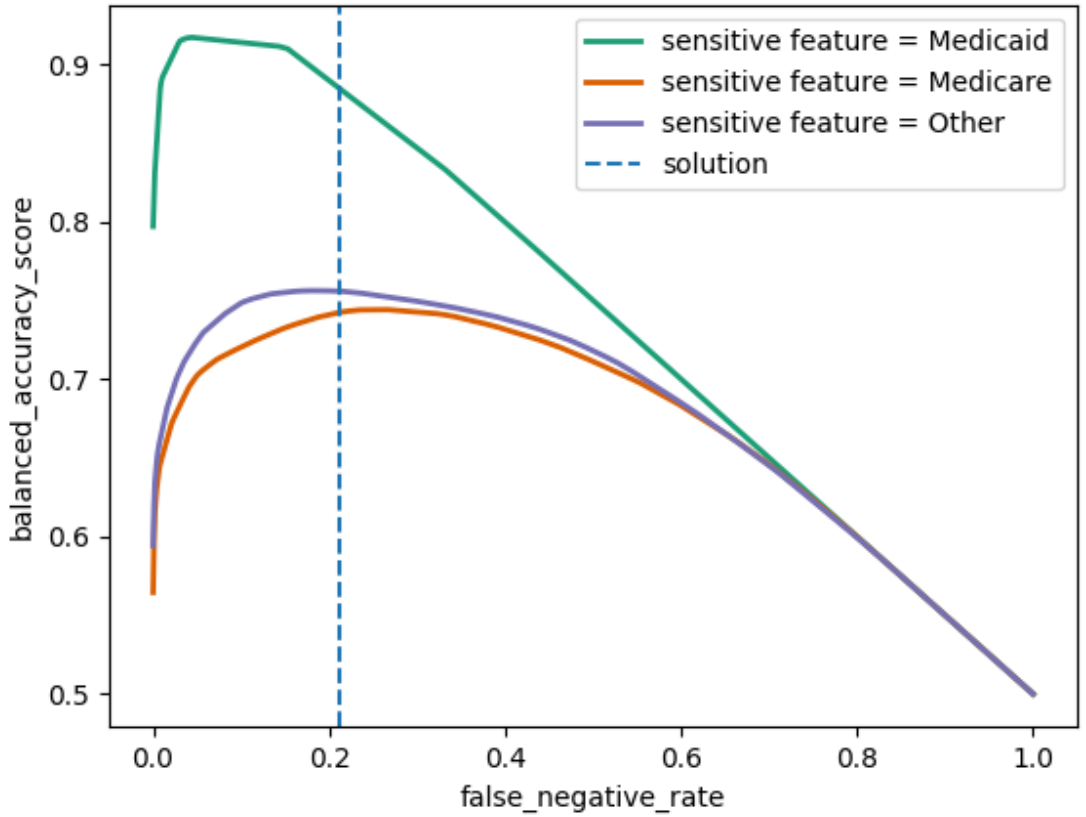


Figure 24: Threshold optimisation based on the sensitive attribute: insurance type. Balanced accuracy is optimised for this subgroup.

finetuned for downstream tasks. Whilst originally trained on a forecasting task, this could be adapted for length-of-stay prediction.

5.3 Early, joint and late fusion

We compared two feature-level early fusion techniques, and found that simple concatenation was sufficient for fairer and more accurate models. Future work comparing early, joint and late fusion methods to one another could allow for the development of more effective techniques. Whilst existing literature review of Huang and colleagues [26, 2] has showed that many studies utilise early fusion, this is often dependent on the range and type of modalities included. For example, when combining data from data with large differences in structure and dimensionality such as imaging, text and tabular features, late fusion can be useful as representations can be formed for each modality whilst optimising for the downstream task. On the other hand, it may be advantageous to use a joint or early framework, to extract more generalisable representations of the data [37]. This is typically the approach used for building foundation model frameworks [51], where high-level representations of the data are required that may not be tailored to a particular predictive task. Therefore, investigating a range of case studies, or different combinations of data modalities will help to shed light on which level and type of fusion are most effective.

5.4 Explaining failure modes with multimodal data

For our case study, we applied fusion methods to combine time-series and static tabular data. However, the use of an additional, unseen modality could be explored to see whether this can explain model behaviour. For example, for a model trained on imaging and tabular information, could we examine the associated radiological reports to explain

why a model produced an incorrect prediction on their scan? Similarly, in the case of time-series data, could we gain insights into failure cases by examining a patients' medical notes. Whilst the modelling in this case may not incorporate the addditional modality, this end-to-end pipeline still leverages multiple sources of information.

5.5 Generating modality-level explanations and causality

Moreover, in terms of explainability we have focused on the influence and importance of features. In order to leverage time-dependence, we compare the important features using different hidden states e.g., $t = 0$ versus $t = 5$. However, since the dynamic features should little variation, the influence on these features prediction remained stable over time. Instead, future work should shift focus on to the interplay between the modalities themselves. For example, when is time-series data more useful over text data?

Post-hoc explanations only reveal part of the picture. In order to generate more informative insights, causality could be investigated further as recent research efforts have begun to move towards building inherently interpretable models [39]. Similarly, casual explanations aim to go beyond ranking features by their influence, to generate a description of their interaction with one another specifically, the presence any directed cause and effect relationships.

6 Conclusion

This project aimed to decipher the impact of multimodal modelling choices for understanding explainability and fairness in a healthcare context. We focus on a case study that utilises electronic health data from the MIMIC-IV dataset, comprised of demographic information and time-varying dynamic features from laboratory events and vital signs. We found an 11% improvement in performance (balanced accuracy) by incorporating time-varying features, and a further 14% improvement when a using an LSTM to model the time-series data in a fusion model. This highlights the importance of combining multiple data sources to leverage complex information and relationships between features. Different fusion strategies can be used to combine data, and we found concatenation sufficient based on predictive accuracy and fairness metrics such as equalised odds and demographic parity. There was little difference between a multi-adaptation gate (MAG) and feature concatenation in terms of performance, however we observed a substantial improvement in training convergence depending on the primary modality (MAG-TS).

SHAP explanations were generated for the various models and were consistent influential features. This is a reassuring finding, since explanation methods are often criticised due to their lack of robustness to model parameters and architectures. However, further work is needed to gain insights into the interplay between different modalities in a multimodal context. The MIMIC-IV dataset includes several protected characteristics such as sex, age, insurance type and marital status. We found a relationship between age and insurance type and the target variable, LOS. This motivated further exploration into the impact of data bias on model explanations and fairness. Initial results suggest that fusion models are more robust to data bias, with all three outperforming their unimodal counterparts based on accuracy and false negative rates within different subpopulations. The fusion model based on feature concatenation generated more fair predictions across the different insurance type categories.

We propose several opportunities to extend this work further, such as by incorporating text as an additional modality. For example, we highlighted the presence of missingness (not-at-random) in the dataset which can cause undesirable correlations between the target variable and specific subgroups. Ablation studies may be useful for systematically evaluating and quantifying the impact of bias such as this on model performance, explainability and fairness. This will likely require access to synthetic data generation models in order to control the underlying feature distributions. Additionally, we identified the presence of bias due to class imbalance, with a focus on insurance type. We demonstrate how this can be mitigated post-hoc with techniques such as threshold optimisation. Whilst this was a simple approach, in-training bias mitigation could offer a more effective strategy for generating fairer models and should be investigated further in the context of multimodal analysis.

7 Appendix

7.1 Out-of-scope ideas

Due to the exploratory nature of this project, the first step involved considering different case studies or applications of multimodal modelling in healthcare. Here, I will detail some alternative project ideas that were discussed which may inform future investigation. In this work, we chose to constrain our ideas to applications relevant to population health research, and consider the data and modalities that would be available and accessible to a body such as NHS England. This included shifting focus away from diagnostic tasks to applications with a broader impact on healthcare services such as predicting readmission. We also decided to avoid use cases that involved the use of imaging data, both due to time and computational constraints, quality differences and the current difficulty accessing imaging data across NHS England.

However, there are a number of out-of-scope case studies that were discussed. These are described in Table 5, alongside potential limitations and overall decision based on the project context:

Task	Modalities	Positives	Drawbacks
Visual-question answering	Imaging, text	Interesting modelling problem, leveraging NLP advancements	Requires access to open-access labelled image-text pairs
In-hospital mortality	EHR, text	Impact on management of servives	Has been explored in previous projects
Forecasting	EHR	Non-trivial to extend to multimodal	Limited publicly available longitudinal data
Readmission	EHR, text	Can be useful for resource management	Retrospective data
Length of stay	EHR, imaging, text	Scope to expand to multiple modalities, useful for resource management and triage	Can be difficult to obtain high accuracy

Table 5: Other potential use-cases to explore multimodal fairness and explainability.

- **Dealing with missingness:** What do we do when certain modalities are missing? How does this affect the resulting model explanations? Can we still evaluate bias? Can these scenarios help aid our understanding of which modalities are most useful?
- **Human-in-the-loop:** Incorporating a feedback loop to improve model representation learning using ranked explanations. See paper on image-generation from text inputs.
- **Less common modalities:** Explainability for omics data, speech analysis or time series data/signals e.g., ECG.
- **Explanation validation:** Using an annotated imaging dataset such as CheXmask to validate model explanations.
- **Generative approaches:** Examples where explanations are generated by the model e.g., prototypes, GANs (synthetic images) as seen in this work which also incorporates explainability into the loss/training.
- **Recommender systems:** Triage support or alert system based on longitudinal modelling / forecasting which can either predict whether an additional image is required and when?
- **Unknown bias identification:** Mechanism to identify unknown sources of bias. See this paper which explores this idea.
- **Human-computer interaction exploration:** A deep drive into multimodal explanations e.g., saliency + text based and how this should best be presented in a clinical context. Refer to this review of MM XAI approaches.
- **Video VQA:** A subset of medical VQA in which questions are asked based on a video e.g., instruction video on how to check a pulse.
- **Building a VQA database from clinical reports:** utilising the methods outlined in other papers to generate an NHS-specific VQA database.
- **Audio e.g., Ambient Voice:** Exploring bias and fairness for audio transcription technology.

7.2 Other publicly-available healthcare datasets

Multimodal data describes the combination of a range of data types such as tabulated electronic health records, which may include static information about a patient and also time-series data from vital sign monitoring, as well as imaging (e.g., chest x-rays), genetic data (tabular, high-dimensional), audio or text (clinical notes, histopathological reports). Each source of data provides a different perspective on an individual's healthcare status and many studies have shown that performance improves when data is combined in a unified framework that is able to exploit inter-modality relationships to capture complex interactions and patterns (cite examples).

This project is particularly interested in multimodal approaches within healthcare, therefore in order to demonstrate and investigate the available techniques, we will need access to datasets comprised of different data modalities. These datasets will inform the use-cases (tasks) that we can consider for the project. A list of PhysioNet challenges can be found here. This review highlights some popular datasets for multimodal learning in medical imaging. In this project, we focus more on the use of EHR data for addressing the use of AI for predicting readmission, mortality or generation of clinical reports. This provides an opportunity to utilise the availability of multi-modal databases, which typically are paired with clinical free-text reports (see Section . Some useful datasets for exploring this case study are:

1. INSPECT: A longitudinal dataset comprised of imaging, EHR, clinical reports with associated diagnosis and prognosis labels across $\sim$20k patients (unavailable at time of writing).
2. MIMIC-IV: A large consortium of imaging, clinical reports and EHR datasets from a single hospital in the US (cross-sectional). MIMIC-IV introduces data from the emergency department (ED) seperate from hospital and ICU ward stays.
3. eICU: A collection of EHR data from several hospitals in the US.
4. TxtRayAlign: A generative model which produces radiological reports based on chest x-rays. Could be used to build a synthetic text dataset that could be adapted to contain inherent biases in order to explore fairness (see previous internship).

References

[1] *5.2 Factors affecting confidence in artificial intelligence (AI) during clinical reasoning and decision making (CRDM) | Workforce, training and education | NHS England — digital-transformation.hee.nhs.uk*. https://digital-transformation.hee.nhs.uk/building-a-digital-workforce/dart-ed/horizon-scanning/understanding-healthcare-workers-confidence-in-ai/chapter-5-clinical-use/factors-affecting-confidence-in-ai-during-crdm. [Accessed 21-07-2024].

[2] Julián N. Acosta et al. "Multimodal biomedical AI". In: *Nature Medicine* 28.9 (Sept. 2022), pp. 1773–1784. ISSN: 1546-170X. DOI: 10.1038/s41591-022-01981-2. URL: http://dx.doi.org/10.1038/s41591-022-01981-2.

[3] Philip Adler et al. "Auditing black-box models for indirect influence". In: *Knowledge and Information Systems* 54.1 (Oct. 2017), pp. 95–122. ISSN: 0219-3116. DOI: 10.1007/s10115-017-1116-3. URL: http://dx.doi.org/10.1007/s10115-017-1116-3.

[4] Ana Lawry Aguila et al. "Multi-view-AE: A Python package for multi-view autoencoder models". In: *Journal of Open Source Software* 8.85 (2023), p. 5093. ISSN: 2475-9066. DOI: 10.21105/joss.05093. URL: http://dx.doi.org/10.21105/joss.05093.

[5] *AI Blindspot: A Discovery Process for preventing, detecting, and mitigating bias in AI systems — aiblindspot.media.mit.edu*. https://aiblindspot.media.mit.edu. [Accessed 04-08-2024].

[6] Nurbanu Aksoy et al. "Beyond images: an integrative multi-modal approach to chest x-ray report generation". In: *Frontiers in Radiology* 4 (Feb. 2024). ISSN: 2673-8740. DOI: 10.3389/fradi.2024.1339612. URL: http://dx.doi.org/10.3389/fradi.2024.1339612.

[7] Emily Alsentzer et al. *Publicly Available Clinical BERT Embeddings*. 2019. DOI: 10.48550/ARXIV.1904.03323. URL: https://arxiv.org/abs/1904.03323.

[8] anai.io. *5 Prominent Pillars Of Explainable AI -ANAI anai.io*. [Accessed 05-07-2024]. 2022.

[9] Hyojin Bahng et al. "Learning De-biased Representations with Biased Representations". In: *CoRR* abs/1910.02806 (2019). arXiv: 1910.02806. URL: http://arxiv.org/abs/1910.02806.

[10] Alejandro Barredo Arrieta et al. "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI". In: *Information Fusion* 58 (June 2020), pp. 82–115. ISSN: 1566-2535. DOI: 10.1016/j.inffus.2019.12.012. URL: http://dx.doi.org/10.1016/j.inffus.2019.12.012.

[11] *basics - Bias and Fairness in AI — bias-and-fairness-in-ai-systems.de*. https://bias-and-fairness-in-ai-systems.de/en/basics/. [Accessed 04-08-2024].

[12] Nadine Bienefeld et al. "Solving the explainable AI conundrum by bridging clinicians' needs and developers' goals". In: *NPJ Digit. Med.* 6.1 (2023), p. 94.

[13] Alexander Binder et al. "Layer-wise relevance propagation for neural networks with local renormalization layers". In: *arXiv* (2016).

[14] Alessandro Castelnovo et al. "A clarification of the nuances in the fairness metrics landscape". In: *Scientific Reports* 12.1 (Mar. 2022). ISSN: 2045-2322. DOI: 10.1038/s41598-022-07939-1. URL: http://dx.doi.org/10.1038/s41598-022-07939-1.

[15] Nitesh V. Chawla et al. "SMOTE: synthetic minority over-sampling technique". In: *J. Artif. Int. Res.* 16.1 (June 2002), pp. 321–357. ISSN: 1076-9757.

[16] IBM Data and AI Team. *Shedding light on AI bias with real world examples - IBM Blog — ibm.com*. https://www.ibm.com/blog/shedding-light-on-ai-bias-with-real-world-examples/. [Accessed 05-07-2024]. 2023.

[17] Encord. *How To Mitigate Bias in Machine Learning Models — encord.com*. https://encord.com/blog/reducing-bias-machine-learning/#:~:text=Pre%2Dprocessing%20and%20Post%2Dprocessing,balance%20predictions%20across%20different%20groups.. [Accessed 04-08-2024].

[18] Emilio Ferrara. "Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies". In: *Sci* 6.1 (2023), p. 3. ISSN: 2413-4155. DOI: 10.3390/sci6010003. URL: http://dx.doi.org/10.3390/sci6010003.

[19] Nazanin Fouladgar and Kary Främling. "A novel LSTM for multivariate time series with massive missingness". en. In: *Sensors (Basel)* 20.10 (May 2020), p. 2832.

[20] Marzyeh Ghassemi, Luke Oakden-Rayner, and Andrew L Beam. "The false hope of current approaches to explainable artificial intelligence in health care". In: *The Lancet Digital Health* 3.11 (Nov. 2021), e745–e750. ISSN: 2589-7500. DOI: 10.1016/s2589-7500(21)00208-9. URL: http://dx.doi.org/10.1016/S2589-7500(21)00208-9.

[21] Moritz Hardt, Eric Price, and Nathan Srebro. "Equality of opportunity in supervised learning". In: *Proceedings of the 30th International Conference on Neural Information Processing Systems*. NIPS'16. Barcelona, Spain: Curran Associates Inc., 2016, pp. 3323–3331. ISBN: 9781510838819.

[22] Hrayr Harutyunyan et al. "Multitask learning and benchmarking with clinical time series data". en. In: *Sci. Data* 6.1 (June 2019), p. 96.

[23] Nasir Hayat, Krzysztof J. Geras, and Farah E. Shamout. *MedFuse: Multi-modal fusion with clinical time-series data and chest X-ray images*. 2022. DOI: 10.48550/ARXIV.2207.07027. URL: https://arxiv.org/abs/2207.07027.

[24] Sepp Hochreiter and Jürgen Schmidhuber. "Long Short-Term Memory". In: *Neural Computation* 9.8 (Nov. 1997), pp. 1735–1780. ISSN: 1530-888X. DOI: 10.1162/neco.1997.9.8.1735. URL: http://dx.doi.org/10.1162/neco.1997.9.8.1735.

[25] Holistic AI. *Holistic AI - AI Governance Platform - holisticai.com*. https://www.holisticai.com. [Accessed 04-08-2024].

[26] Shih-Cheng Huang et al. "Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines". In: *npj Digital Medicine* 3.1 (Oct. 2020). ISSN: 2398-6352. DOI: 10.1038/s41746-020-00341-z. URL: http://dx.doi.org/10.1038/s41746-020-00341-z.

[27] Toshihiro Kamishima et al. "Fairness-Aware Classifier with Prejudice Remover Regularizer". In: *Machine Learning and Knowledge Discovery in Databases*. Ed. by Peter A. Flach, Tijl De Bie, and Nello Cristianini. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 35–50. ISBN: 978-3-642-33486-3.

[28] Changwoo Kim et al. "Fairness-Aware Multimodal Learning in Automatic Video Interview Assessment". In: *IEEE Access* 11 (2023), pp. 122677–122693. DOI: 10.1109/ACCESS.2023.3325891.

[29] Felix Krones et al. "Review of multimodal machine learning approaches in healthcare". In: *ArXiv* abs/2402.02460 (2024). URL: https://api.semanticscholar.org/CorpusID:267412288.

[30] Yikuan Li et al. "BEHRT: Transformer for Electronic Health Records". In: *Scientific Reports* 10.1 (Apr. 2020). ISSN: 2045-2322. DOI: 10.1038/s41598-020-62922-y. URL: http://dx.doi.org/10.1038/s41598-020-62922-y.

[31] Yurui Li, Mingjing Du, and Sheng He. "Attention-Based Sequence-to-Sequence Model for Time Series Imputation". In: *Entropy* 24.12 (Dec. 2022), p. 1798. ISSN: 1099-4300. DOI: 10.3390/e24121798. URL: http://dx.doi.org/10.3390/e24121798.

[32] Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. "Foundations & trends in multimodal machine learning: Principles, challenges, and open questions". en. In: *ACM Comput. Surv.* 56.10 (2024), pp. 1–42.

[33] Zachary C. Lipton, David C. Kale, and Randall Wetzel. *Modeling Missing Data in Clinical Time Series with RNNs*. 2016. arXiv: 1606.04130 [cs.LG]. URL: https://arxiv.org/abs/1606.04130.

[34] Scott M Lundberg and Su-In Lee. "A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon et al. Curran Associates, Inc., 2017, pp. 4765–4774. URL: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

[35] Moxuan Ma et al. "Research on multimodal fusion of temporal electronic medical records". en. In: *Bioengineering (Basel)* 11.1 (Jan. 2024).

[36] Chuizheng Meng et al. "Interpretability and fairness evaluation of deep learning models on MIMIC-IV dataset". en. In: *Sci. Rep.* 12.1 (May 2022), p. 7166.

[37] Farida Mohsen et al. "Artificial intelligence-based methods for fusion of electronic health records and imaging data". In: *Scientific Reports* 12.1 (Oct. 2022). ISSN: 2045-2322. DOI: 10.1038/s41598-022-22514-4. URL: http://dx.doi.org/10.1038/s41598-022-22514-4.

[38] Christoph Molnar. *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*. 2nd ed. GitHub, 2022. URL: https://christophm.github.io/interpretable-ml-book.

[39] Raha Moraffah et al. *Causal Interpretability for Machine Learning – Problems, Methods and Evaluation*. 2020. arXiv: 2003.03934 [cs.LG]. URL: https://arxiv.org/abs/2003.03934.

[40] Xiangdong Pei et al. "A Review of the Application of Multi-modal Deep Learning in Medicine: Bibliometrics and Future Directions". In: *International Journal of Computational Intelligence Systems* 16.1 (2023), p. 44. ISSN: 1875-6883. DOI: 10.1007/s44196-023-00225-6. URL: https://doi.org/10.1007/s44196-023-00225-6.

[41] Geoff Pleiss et al. "On Fairness and Calibration". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/b8b9c74ac526fffbeb2d39ab038d1cd7-Paper.pdf.

[42] Shangran Qiu et al. "Multimodal deep learning for Alzheimer's disease dementia assessment". In: *Nature Communications* 13.1 (June 2022). ISSN: 2041-1723. DOI: 10.1038/s41467-022-31037-5. URL: http://dx.doi.org/10.1038/s41467-022-31037-5.

[43] Wasifur Rahman et al. *Integrating Multimodal Information in Large Pretrained Transformers*. 2019. DOI: 10.48550/ARXIV.1908.05787. URL: https://arxiv.org/abs/1908.05787.

[44] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ""Why Should I Trust You?": Explaining the Predictions of Any Classifier". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*. 2016, pp. 1135–1144.

[45] María Agustina Ricci Lara, Rodrigo Echeveste, and Enzo Ferrante. "Addressing fairness in artificial intelligence for medical imaging". en. In: *Nat. Commun.* 13.1 (Aug. 2022), p. 4581.

[46] Waddah Saeed and Christian Omlin. "Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities". In: *Knowledge-Based Systems* 263 (2023), p. 110273. ISSN: 0950-7051. DOI: 10.1016/j.knosys.2023.110273. URL: http://dx.doi.org/10.1016/j.knosys.2023.110273.

[47] Andrew D Selbst and Julia Powles. "Meaningful information and the right to explanation". In: *International Data Privacy Law* 7.4 (Nov. 2017), pp. 233–242. ISSN: 2044-4001. DOI: 10.1093/idpl/ipx022. URL: http://dx.doi.org/10.1093/idpl/ipx022.

[48] Ramprasaath R Selvaraju et al. "Grad-CAM: Visual explanations from deep networks via gradient-based localization". en. In: *Int. J. Comput. Vis.* 128.2 (Feb. 2020), pp. 336–359.

[49] Luis R Soenksen et al. "Integrated multimodal artificial intelligence framework for healthcare applications". en. In: *NPJ Digit. Med.* 5.1 (Sept. 2022), p. 149.

[50] Ethan Steinberg et al. "MOTOR: A Time-to-Event Foundation Model For Structured Medical Records". In: *The Twelfth International Conference on Learning Representations*. 2024. URL: https://openreview.net/forum?id=NialiwI2V6.

[51] Chameleon Team. *Chameleon: Mixed-Modal Early-Fusion Foundation Models*. 2024. arXiv: 2405.09818 [cs.CL]. URL: https://arxiv.org/abs/2405.09818.

[52] Florence Townend, Patrick J. Roddy, and Philipp Goebl. *florencejt/fusilli: Fusilli v1.0.0*. 2023. DOI: 10.5281/ZENODO.10228564. URL: https://zenodo.org/doi/10.5281/zenodo.10228564.

[53] Ashish Vaswani et al. *Attention Is All You Need*. 2017. DOI: 10.48550/ARXIV.1706.03762. URL: https://arxiv.org/abs/1706.03762.

[54] Jason Walonoski et al. "Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record". In: *Journal of the American Medical Informatics Association* 25.3 (Aug. 2017), pp. 230–238. ISSN: 1527-974X. DOI: 10.1093/jamia/ocx079. eprint: https://academic.oup.com/jamia/article-pdf/25/3/230/34150150/ocx079.pdf. URL: https://doi.org/10.1093/jamia/ocx079.

[55] Yuanlong Wang, Changchang Yin, and Ping Zhang. "Multimodal risk prediction with physiological signals, medical images and clinical notes". en. In: *Heliyon* 10.5 (Mar. 2024), e26772.

[56] Zhenbang Wu et al. "Multimodal Patient Representation Learning with Missing Modalities and Labels". In: *The Twelfth International Conference on Learning Representations*. 2024. URL: https://openreview.net/forum?id=Je5SHCKpPa.

[57] Bo Yang and Lijun Wu. "How to leverage multimodal EHR data for better medical predictions?" In: *arXiv* (Oct. 2021). eprint: 2110.15763 (cs.CL).

[58] Zhichao Yang et al. "TransformEHR: transformer-based encoder-decoder generative model to enhance prediction of disease outcomes using electronic health records". In: *Nature Communications* 14.1 (Nov. 2023). ISSN: 2041-1723. DOI: 10.1038/s41467-023-43715-z. URL: http://dx.doi.org/10.1038/s41467-023-43715-z.

[59] Rich Zemel et al. "Learning Fair Representations". In: *Proceedings of the 30th International Conference on Machine Learning*. Ed. by Sanjoy Dasgupta and David McAllester. Vol. 28. Proceedings of Machine Learning Research 3. Atlanta, Georgia, USA: PMLR, June 2013, pp. 325–333. URL: https://proceedings.mlr.press/v28/zemel13.html.

[60] Xianbing Zhao et al. "MAG+: An Extended Multimodal Adaptation Gate for Multimodal Sentiment Analysis". In: *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2022, pp. 4753–4757. DOI: 10.1109/ICASSP43922.2022.9746536.