applied sciences

MDPI

Article

Use of Attention Maps to Enrich Discriminability in Deep Learning Prediction Models Using Longitudinal Data from Electronic Health Records

Lucía A. Carrasco-Ribelles [1,2,3], Margarita Cabrera-Bean [2], Jose Llanes-Jurado [4] and Concepción Violán [3,5,6,7,*]

[1] Institut Universitari d'Investigació en Atenció Primària Jordi Gol (IDIAP Jordi Gol), 08007 Barcelona, Spain; lcarrasco@idiapjgol.info
[2] Department of Signal Theory and Communications, Universitat Politècnica de Catalunya (UPC), 08034 Barcelona, Spain; marga.cabrera@upc.edu
[3] Unitat de Suport a la Recerca Metropolitana Nord, Institut Universitari d'Investigació en Atenció Primària Jordi Gol (IDIAP Jordi Gol), 08303 Mataró, Spain
[4] Instituto Universitario de Investigación en Tecnología Centrada en el Ser Humano, Universitat Politècnica de València, 46022 Valencia, Spain
[5] Direcció d'Atenció Primària Metropolitana Nord Institut Català de Salut, 08915 Badalona, Spain
[6] Fundació Institut d'Investigació en Ciències de la Salut Germans Trias i Pujol (IGTP), 08916 Badalona, Spain
[7] Department of Medicine, Universitat Autònoma de Barcelona, 08913 Cerdanyola del Vallès, Spain
* Correspondence: cviolanf.mn.ics@gencat.cat

Featured Application: A better discrimination in a prediction model does not imply a better interpretability. In healthcare, where transparency is crucial, both discriminability and interpretability should be checked before stating that a new model is better.

Abstract: Background: In predictive modelling, particularly in fields such as healthcare, the importance of understanding the model's behaviour rivals, if not surpasses, that of discriminability. To this end, attention mechanisms have been included in deep learning models for years. However, when comparing different models, the one with the best discriminability is usually chosen without considering the clinical plausibility of their predictions. Objective: In this work several attention-based deep learning architectures with increasing degrees of complexity were designed and compared aiming to study the balance between discriminability and plausibility with architecture complexity when working with longitudinal data from Electronic Health Records (EHRs). Methods: We developed four deep learning-based architectures with attention mechanisms that were progressively more complex to handle longitudinal data from EHRs. We evaluated their discriminability and resulting attention maps and compared them amongst architectures and different input processing approaches. We trained them on 10 years of data from EHRs from Catalonia (Spain) and evaluated them using a 5-fold cross-validation to predict 1-year all-cause mortality in a subsample of 500,000 people over 65 years of age. Results: Generally, the simplest architectures led to the best overall discriminability, slightly decreasing with complexity by up to 8.7%. However, the attention maps resulting from the simpler architectures were less informative and less clinically plausible compared to those from more complex architectures. Moreover, the latter could give attention weights both in the time and feature domains. Conclusions: Our results suggest that discriminability and more informative and clinically plausible attention maps do not always go together. Given the preferences within the healthcare field for enhanced explainability, establishing a balance with discriminability is imperative.

check for updates

Academic Editors: Han Liu, Hongyang Chen and Xiaotong Zhang

Received: 22 November 2024
Revised: 21 December 2024
Accepted: 22 December 2024
Published: 27 December 2024

Citation: Carrasco-Ribelles, L.A.; Cabrera-Bean, M.; Llanes-Jurado, J.; Violán, C. Use of Attention Maps to Enrich Discriminability in Deep Learning Prediction Models Using Longitudinal Data from Electronic Health Records. *Appl. Sci.* **2025**, *15*, 146. https://doi.org/10.3390/app15010146

Keywords: attention mechanism; clinical plausibility; discriminability; electronic health record; recurrent neural network; longitudinal data

1. Introduction

The rise in the use of real-world data and electronic health records (EHRs) has made large-scale longitudinal health studies [1,2] more cost-effective. Longitudinal data provide a more accurate account of patients' health history over time, facilitate the study of temporal associations and patterns [3], and enhance prediction model performance [4]. Unlike classical statistical methods, machine learning techniques are ideal for harnessing the potential of EHRs for clinical predictions due to their ability to handle numerous variables with non-linear relationships [5] and challenges associated with EHRs like missing values, heterogeneity, and temporal sparsity.

Healthcare workers usually prefer transparent models over "black-box" ones, seeking clarity in informed decision-making [6]. While machine learning-based models lack natural explainability, recent approaches aim to identify feature contribution to the outcome [7], improving interpretability [6]: attention mechanisms [8–10], gradient-based methods [11], and model-agnostic techniques (i.e., LIME [12], SHAP [13]). These contributions can be calculated in different ways (e.g., feature or time dimension) to obtain more complete information [14–16]. Attention mechanisms can also potentially enhance model performance [17]. Some machine learning techniques can analyse EHRs with attention mechanisms to support decision-making [5], most being based on sequential deep learning architectures [14,16,18–20]. Benchmarking studies usually assess only discriminability [14,21,22], but attention weights' clinical relevance also requires scrutiny, ensuring meaningful explanations align with clinical insights. As noted in a previous systematic review [18], there is a lack of guidelines in how to run benchmarking studies on prediction models in healthcare, hampering comparability between models, especially considering interpretability. Recently developed reporting guidelines for prediction models in healthcare with AI such as TRIPOD-AI [23] do not define how to perform these comparisons either.

In this study, several attention-based deep learning architectures handling longitudinal EHR data were designed with increasing degrees of complexity and compared, aiming to test whether improved discriminability necessarily implies improved interpretability. To do so, we designed multiple architectures with attention mechanisms at different levels and analysed their impact not only in terms of discriminability but also in the distribution of attention weights. The evaluation was performed on all-cause mortality with real-world EHRs.

2. Related Work

Sequential models like recurrent neural networks (RNNs) handle longitudinal data, assuming registers are evenly time-spaced [5]. However, real-world EHRs could be irregular. Some methods like RETAIN include an extra feature indicating time intervals between visits [16,24,25]. Others aggregate data within time windows while maintaining certain temporal evolution. For instance, Patient2Vec requires a consistent number of time periods across patients [14], unlike ZiMM [19]. Training on batches of patients with equal visit counts [26,27] as well as zero padding and masking are also common. A systematic review up to 2022 found that models using longitudinal EHR data usually involved less than three hidden layers [18]. While 61.7% used architectures based on RNNs only, 27.2% used combinations of layers like convolutional neural networks (CNNs) or graph neural networks (GNNs), with or without

RNNs [18]. Lately, temporal CNNs (TCNNs) and transformers have shown similar or better results than RNNs when working with longitudinal tabular data [28,29].

Attention mechanisms are a component that can be included in deep learning architectures capable of providing an intuition of the contribution of each input to the outcome. They were first introduced in natural language processing, specifically within sequence-to-sequence tasks like machine translation [9] and then expanded to other domains. Extended taxonomy reviews have been presented [30,31]. In healthcare, many studies have included them, aiming to increase their transparency [18], often in medical imaging, where attention maps help to locate diseases [32,33], and less often with longitudinal tabular data from EHRs. In these cases, in addition to transparency, attentional mechanisms also help to handle long temporal sequences more successfully. Choi et al. (2016) introduced RETAIN, a model with a parallel attention mechanism for both variable and time dimension predicting heart failure using embedded features rather than the immediate input features [16]. Patient2Vec by Zhang et al. (2018) employed a hierarchical attention mechanism that learned dedicated weights for the clinical events in each visit and aggregated weights for the visits to predict the next diagnosis [14]. Similarly, Kabeshova et al. (2020) proposed ZiMM, combining self-attention with recurrent layers into an encoder–decoder architecture to predict long-term and blurry relapses [19].

While the role of attention mechanisms in explainability is currently debated [34–36] due to the lack of a consistent definition of explainability, their contribution to transparency is not discussed [35]. Sen et al. performed a study to assess how good attention maps were based on the consensus with human explainability [37], following the recommendations from Riedl on human-centred artificial intelligence to provide a good explanation such as one that is plausible [38]. They defined some metrics to measure the consensus, such as the overlap between humans and attention mechanisms but recognized that the subjectivity of humans themselves could make it difficult to define these metrics.

3. Materials and Methods

3.1. Proposed Architectures

Figure 1 shows the different approaches that were designed and compared in this work, starting with an initial recurrent layer-based architecture without attention mechanisms (Approach 0). Then, in Approach 1, we compared four kinds of attention layers reporting the contribution of each time period added to the starting architecture. Approach 2, inspired by RETAIN [16], included a feature domain branch that ran in parallel to that from Approach 1, providing attention weights for both feature and temporal domains (i.e., "domain-specific" attention). Finally, Approach 3, similarly to Patient2Vec [14], considered a "hierarchical" attention architecture obtaining attention weights for the feature domain in each time period and an overall attention weight for the time domain. The kind of attention layer producing the best results in Approach 1 was selected for Approaches 2 and 3, while comparing different ways of processing inputs. Notation and model equations are included in Supplementary Materials.

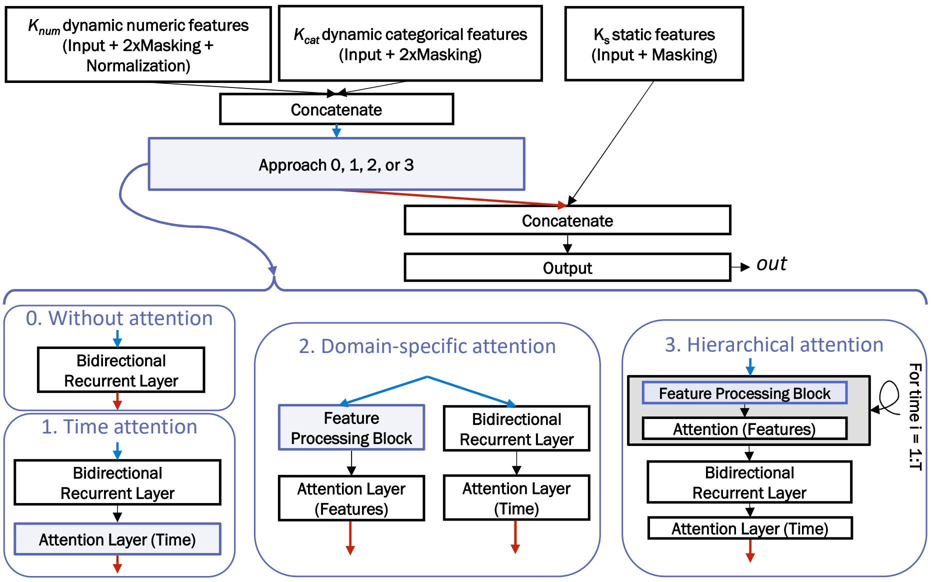
K_{num} dynamic numeric features (Input + 2xMasking + Normalization)
K_{cat} dynamic categorical features (Input + 2xMasking)
K_s static features (Input + Masking)
Concatenate
Approach 0, 1, 2, or 3
Concatenate
Output
out
0. Without attention
Bidirectional Recurrent Layer
1. Time attention
Bidirectional Recurrent Layer
Attention Layer (Time)
2. Domain-specific attention
Feature Processing Block
Bidirectional Recurrent Layer
Attention Layer (Features)
Attention Layer (Time)
3. Hierarchical attention
Feature Processing Block
Attention (Features)
For time i = 1:T
Bidirectional Recurrent Layer
Attention Layer (Time)

Figure 1. General scheme of the architectures that were evaluated. Input data processing and output generation (top and bottom boxes) were common to all the tested architectures. The purple box represents any of the four approaches shown in the bottom boxes. The shadowed boxes identify which parts in each approach compared different options. Coloured arrows indicate the input (blue) and output (red). Notation can be found in the Supplementary Materials.

3.1.1. Input

Each subject's information is recorded in EHR systems through K dynamic variables, which can be numeric (K_{num}) or categorical (K_{cat}), with $K = K_{num} + K_{cat}$; and K_S static variables. Similarly to [14], we temporally aggregated each of the K dynamic variables in 1-year time windows (see Section 3.2, Figure S1) to partially compact the information on the time axis. Then, we implemented a two-level masking process. The first one addressed temporal inconsistency between subjects, ensuring that all subjects had the same length in the temporal axis, while the second addressed the missingness in the original K and K_S features, per patient and year (see Figure 1 and Supplementary Materials for further information).

3.1.2. Baseline Architecture Without Attention (Approach 0)

This baseline approach was based on a single bidirectional RNN ($BiRL$) and did not include an attention mechanism. Its complete formulation and that of the following architectures can be found in Supplementary Materials.

3.1.3. Basic Architecture with Time Attention (Approach 1)

Based on the baseline architecture in Section 3.1.2, an attention layer was included after the $BiRL$ to provide the contribution of each time period t to the outcome through an attention weights vector, $\boldsymbol{\alpha}_T \in \mathbb{R}^T$. We compared four kinds of attention layers, as formulated in Equations (1) [19], (2) [39], (3), and (4).

$$\boldsymbol{\alpha}_T = sf(\mathbf{w}_1^\top \mathbf{H}_T + \mathbf{b}_1); \quad \mathbf{c}_T = \mathbf{H}_T \boldsymbol{\alpha}_T \tag{1}$$

$$\boldsymbol{\alpha}_T = sf(\mathbf{w}_1^\top \tanh(\mathbf{H}_T)); \quad \mathbf{c}_T = \tanh(\mathbf{H}_T \boldsymbol{\alpha}_T) \tag{2}$$

$$\boldsymbol{\alpha}_T = sf(\mathbf{w}_1^\top \mathbf{H}_T + \mathbf{b}_1); \quad \mathbf{c}_T = \tanh(\mathbf{H}_T \boldsymbol{\alpha}_T) \tag{3}$$

$$\boldsymbol{\alpha}_T = sf(\mathbf{w}_1^\top \tanh(\mathbf{H}_T)); \quad \mathbf{c}_T = \mathbf{H}_T \boldsymbol{\alpha}_T \tag{4}$$

sf denotes the well-known softmax normalization function, and tanh is the hyperbolic tangent. As detailed in Supplementary Materials, in Equations (1) to (4), matrix $\mathbf{H}_T \in \mathbb{R}^{2U \times T}$, where U is the number of neurons of the $BiRL$, contains the BiRL output sequence (forward and backward), and vector $\mathbf{c}_T \in \mathbb{R}^{2U}$ is the so-called context vector needed for output prediction. Finally, vectors $\mathbf{w}_1 \in \mathbb{R}^{2U}$ and $\mathbf{b}_1 \in \mathbb{R}^{T}$ are learned during the training process.

3.1.4. Architecture with Domain-Specific Attention (Approach 2)

This domain-specific architecture incorporated a new branch focused on the feature domain that provided attention weights for each feature through vector $\boldsymbol{\alpha}_f$, in parallel to those weights for the time domain, $\boldsymbol{\alpha}_T$. Figure 2 shows the set of input feature processing methods that were compared, and a subsection is devoted in Supplementary Materials for each.

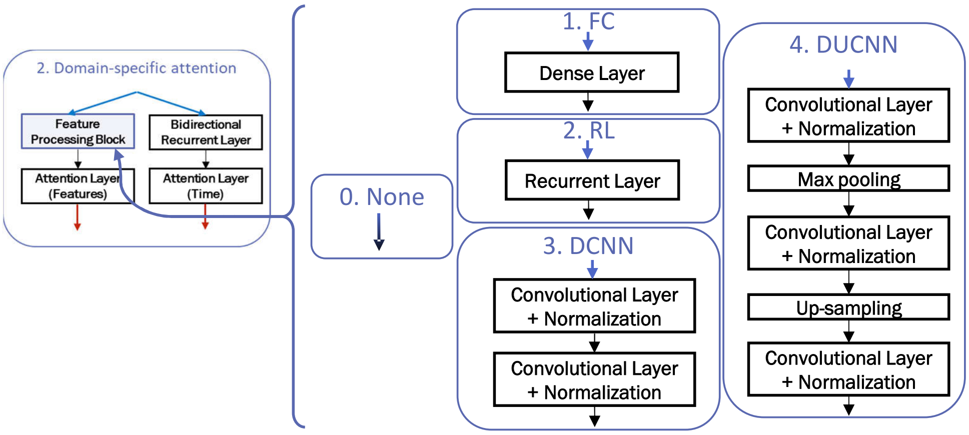
2. Domain-specific attention
Feature Processing Block
Bidirectional Recurrent Layer
Attention Layer (Features)
Attention Layer (Time)
0. None
1. FC
Dense Layer
2. RL
Recurrent Layer
3. DCNN
Convolutional Layer + Normalization
Convolutional Layer + Normalization
4. DUCNN
Convolutional Layer + Normalization
Max pooling
Convolutional Layer + Normalization
Up-sampling
Convolutional Layer + Normalization

Figure 2. Options for processing the feature domain in the domain-specific and hierarchical attention approaches (Approaches 2 and 3, respectively). The blue arrow indicates the input to each block, as in Figure 1. FC: dense, fully connected; RL: recurrent layer; DCNN: down-sampling convolutional neural network; DUCNN: down–up-sampling convolutional neural network.

3.1.5. Architecture with Hierarchical Attention (Approach 3)

This last architecture aimed to provide the contribution of each feature k in each time period t to the outcome prediction, as well as an overall contribution for each t. For this purpose, a hierarchical attention approach was implemented. First, we applied one of the previous feature processing methods (see Section 3.1.4, Figure 2) per time period. The resulting processed vectors in each time period were then concatenated and passed as input to the time-domain block with the same architecture as in Section 3.1.3. As in the previous approach, all the feature processing methods were compared. The complete formulation can be found in Supplementary Materials.

3.2. Evaluation

3.2.1. Study Design, Participants, and Data Source

Data were drawn from SIDIAP, a primary care EHR database from Catalonia (Spain) [40]. SIDIAP includes data on demographics, visits to primary care, diagnoses, laboratory and clinical measurements, and drugs, among others.

A dynamic cohort was drawn between 1 January 2010 and 31 December 2019, including participants either aged ≥65 at baseline or throughout the study. They were followed until death or transfer out of the catchment area. Individuals with no available data, no visits to the General Practitioner (GP) during the study period, or aged 100 years or older in 2010 were excluded. From this sample, a random subsample of 500,000 individuals was selected (Figure S2). Among these participants, 280,330 (56.1%) were female, the initial age was 71.9 ± 7.7 (mean ± standard deviation), and 24.2% died within one year. A more detailed description of the population and the outcome can be found in Supplementary Materials.

The outcome was all-cause mortality at one year. Around 250 features (see Table 1) were extracted from the EHR and aggregated on an annual basis using the mean, if quantitative, or the mode, if qualitative. This left a mean of 7 records (years) (standard deviation: 3.15) per person. The annual aggregation was chosen over a more granular method because the data source was primary care. Visiting the GP several times in a single year is uncommon, especially in younger people, so greater temporal granularity would have increased the number of missing values. The complete list of variables can be found in [41]. Missingness was addressed in this work through masking. Further details can be found in Supplementary Materials.

Table 1. Summary of the clinical variables that were extracted from the EHR per data domain. The complete list can be found on our Github [41].

Clinic Domain	Number of Features
Hospital admissions	9
Clinical measurements (e.g., questionnaires, BMI)	50
Primary care diagnosis, as in [42]	64
Drugs	3
Sociodemographics	4
Smoking habits	4
Visits to primary care	8
Laboratory results	67
Frailty deficits, as in [43]	37

3.2.2. Model Development and Discriminability Evaluation

A 5-fold cross-validation was used to evaluate the performance of each architecture, calculating the following metrics on the validation set: recall, precision, Cohen's Kappa, area under the ROC curve (ROC-AUC), and the precision–recall curve (PR-AUC). Mean and standard error were calculated among the 5 repetitions. As there is currently no accepted quantitative metric for measuring the clinical plausibility of attention maps, a family doctor (CV) on the team assessed it, following Riedl's definition of a good explanation [38].

The model was developed with Tensorflow in Python 3.9. To prevent overfitting, training stopped if the ROC-AUC did not improve more than 0.001 after 10 epochs, and the best weights were restored. Batch normalisation and a rate of 0.1 recurrent dropout were also included. The learning rate was reduced by 0.1 if the ROC-AUC did not change after 5 epochs. Batch size was set to 32, and the maximum number of epochs was 25. The optimizer was Adam [44], with an initial learning rate of 0.001. The number of neurons U was consistent in all the architectures and set to 128. The recurrent layer was a GRU in all cases. The kernel size was 5 for all the convolutional layers. While in Approach 1, four different kinds of attention layers were compared, only the best attention layer in that approach was used for Approaches 2 and 3.

4. Results

4.1. Discriminability Analysis

Table 2 shows the discriminability metrics of all the architectures. In Approaches 0 and 1, overall, discriminability was high, with notable recall (i.e., true positives), precision, and Cohen's Kappa (i.e., balance true positives and negatives), even though including the attention layer (Approach 1) led to varying degrees of decrease in all metrics except recall. The attention layer in (2) was chosen for further experiments due to its slightly higher PR-AUC and minimal decrease in metrics (0.98%) compared to (1) and Approach 0, respectively.

Table 2. Mean value (standard deviation) of the metrics of the different architectures. Approaches 0, 1, 2, and 3 correspond to the architectures proposed in Sections 3.1.2–3.1.5, respectively.

Architecture	Cohen's Kappa	PR-AUC	Precision	Recall	ROC-AUC
Approach 0—Without attention	0.81 (0.003)	0.77 (0.003)	0.88 (0.004)	0.82 (0.003)	0.89 (0.001)
Approach 1—Time attention					
Linear Reg. + no final Tanh (1)	0.79 (0.003)	0.75 (0.004)	0.87 (0.003)	0.82 (0.002)	0.89 (0.001)
No linear Reg. + final Tanh (2)	0.79 (0.003)	0.76 (0.004)	0.87 (0.003)	0.82 (0.002)	0.89 (0.002)
Linear Reg. + final Tanh (3)	0.5 (0.3)	0.53 (0.2)	0.58 (0.3)	0.87 (0.05)	0.79 (0.09)
No linear Reg. + no final Tanh (4)	0.6 (0.3)	0.6 (0.2)	0.68 (0.3)	0.85 (0.05)	0.82 (0.09)
Approach 2—Domain-specific attention					
None	0.7 (0.2)	0.68 (0.2)	0.78 (0.2)	0.83 (0.04)	0.85 (0.07)
Dense	0.7 (0.2)	0.68 (0.2)	0.78 (0.2)	0.83 (0.04)	0.85 (0.07)
GRU	0.7 (0.2)	0.69 (0.2)	0.79 (0.2)	0.83 (0.04)	0.86 (0.08)
Down-sampling CNN	0.7 (0.2)	0.68 (0.2)	0.78 (0.2)	0.83 (0.04)	0.85 (0.07)
Down-Up CNN	0.6 (0.3)	0.61 (0.2)	0.69 (0.3)	0.85 (0.05)	0.82 (0.09)
Approach 3—Hierarchical attention					
None	0.28 (0.03)	0.35 (0.02)	0.39 (0.01)	0.75 (0.2)	0.69 (0.04)
Dense	0.45 (0.09)	0.45 (0.05)	0.57 (0.1)	0.66 (0.1)	0.74 (0.01)
GRU	0.61 (0.3)	0.61 (0.2)	0.69 (0.3)	0.85 (0.04)	0.82 (0.1)
Down-sampling CNN	0.69 (0.2)	0.67 (0.2)	0.77 (0.2)	0.83 (0.05)	0.85 (0.07)
Down-Up CNN	0.61 (0.3)	0.61 (0.2)	0.69 (0.3)	0.85 (0.05)	0.83 (0.09)

Considering the different feature-domain processing methods in Approach 2 (Section 3.1.4), the simplest ones (Dense, GRU, and CNN down-sampling) showed similar, best results. On the other hand, the down–up CNN performed worse in true negatives, with higher recall but lower Cohen's Kappa and precision. The architecture with the best discriminability, i.e., the one processing the features using a GRU, showed a 6.39% decrease in discriminability (averaging the changes in all the metrics) compared to the best from Approach 1, mainly due to decreased performance in predicting true negatives (1-year survival), e.g., the precision decreased by 10.34%.

Approach 3 (Section 3.1.5), employing hierarchical attention, compared the feature-domain processing methods again. The down-sampling CNN proved best in terms of discriminability, though slightly lower than the best of Approach 2 (by 1.60%), Approach 1 (7.85%), and baseline Approach 0 (8.72%). As in Approach 2, the drop in discriminability affected only the detection of true negatives, while recall was maintained or even increased. Using a single FC layer for feature processing decreased model discriminability, likely due to class imbalance and more complex processing methods' ability to predict the outcome of interest.

4.2. Attention Maps' Analysis

While in Approach 0, no attention maps could be extracted, they were available in the rest of the approaches (in Approach 1, exclusively in the time domain). Figure 3 shows each time period's contribution to the outcome across approaches. Simpler input feature processing methods (i.e., only time attention, or those with none or FC layers in Approaches 2 and 3) assigned similar attention weights to all time periods, while those with higher complexity (e.g., CNN, GRU) managed to distribute them, assigning higher weights to those closer to the prediction time. This aligns more with clinical plausibility, especially in older, frail patients, where death often follows abrupt changes.

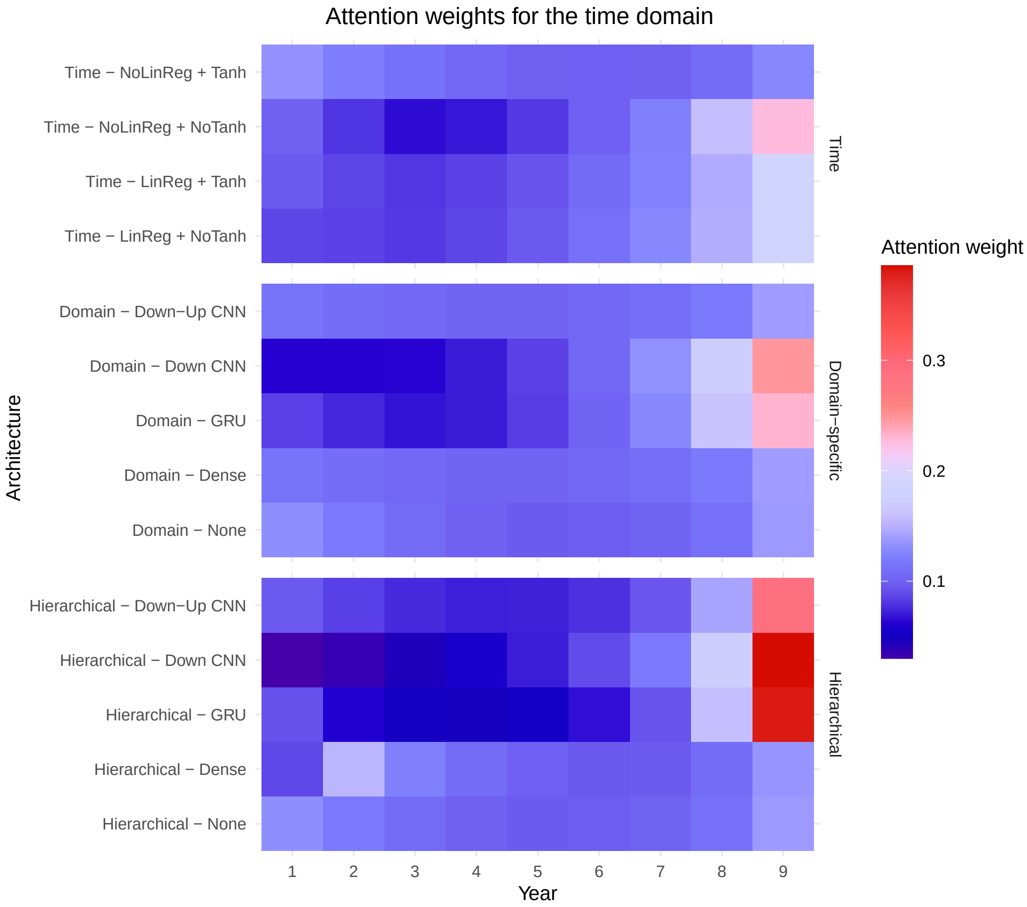

Figure 3. Attention maps on the time domain for all the architectures, by approach. The higher the number on the x-axis, the closer to the moment when the prediction was performed. The attention weights shown here are the results of averaging the attention weights of every patient on the test set.

Figures 4 and 5 show each feature's contribution to the outcome, calculated by domain-specific (Approach 2) and hierarchical attention (Approach 3) methods, respectively. While domain-specific attention offers an overall attention per feature, hierarchical attention provides feature contributions per time period. This last approach, giving different weights to each type of event and different weights to each event in different periods, closely mirrors actual clinical reasoning during patient assessment, generating more informative and realistic attention maps than Approach 2.

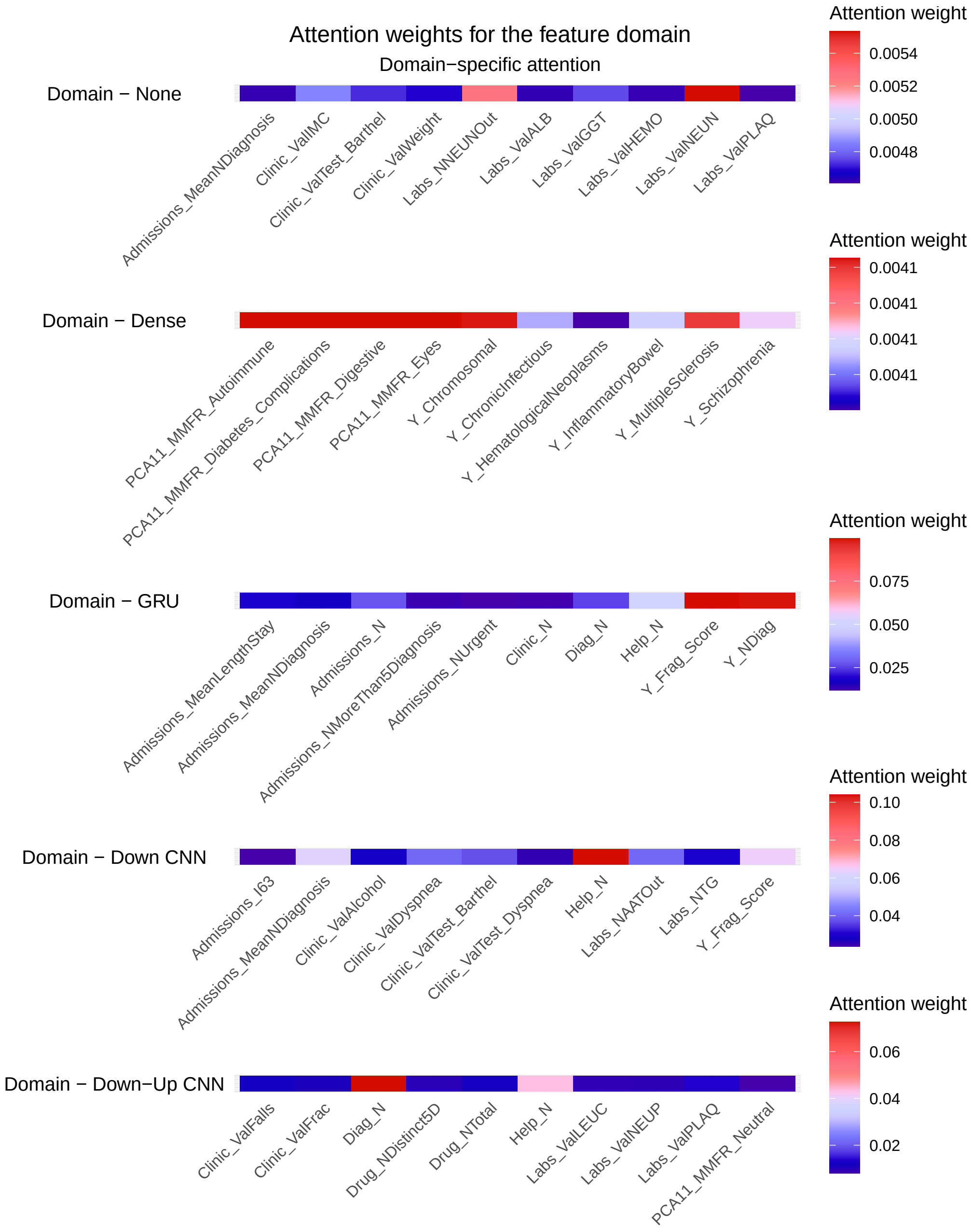

Figure 4. Attention weights on the feature domain for the domain-specific attention architectures (Approach 2). The attention weights shown here are the results of averaging the attention weights of every patient on the test set. The description of each feature can be found on our GitHub [41].

Similarly, simpler input feature processing methods (e.g., FC layer) failed to distribute attention weights among features in both approaches, assigning low weights uniformly. On the other hand, GRU, down-sampling CNN, and down–up CNN achieved a better distribution among features. In Approach 3, the down-sampling CNN failed to distribute attention weights amongst time periods, lacking clinical plausibility, while the GRU or a down–up CNN achieved it in both time and feature domains. However, the GRU's feature processing yielded clinically more plausible attention weights compared to the down–up CNN, prioritizing chronic diagnoses (Y_NDiag), frailty deficits (Y_Frag_Score), and hospital admissions (Admissions_N).

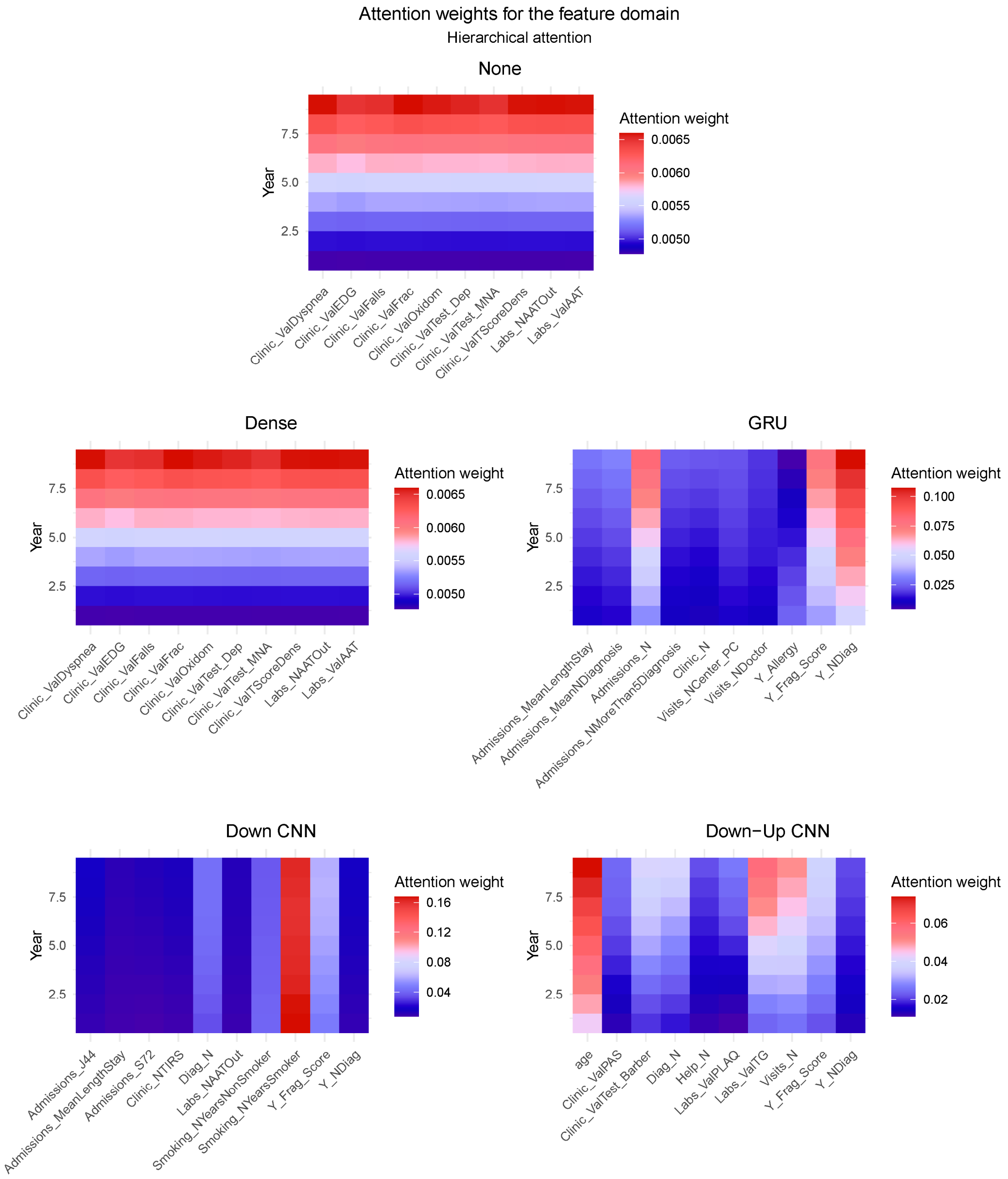

Figure 5. Attention maps on the feature domain for the hierarchical attention architectures (Approach 3). The attention weights shown here are the results of averaging the attention weights of every patient on the test set. The description of each feature can be found on our GitHub [41].

Comparing the top 10 features with the highest attention weights between domain-specific (Approach 2) and hierarchical (Approach 3) attention, there was 0% overlap using the FC layer, 70% with the GRU, 20% with the down-sampling CNN, and 30% with the down–up CNN. Thus, GRU-based feature processing exhibited more consistent attention weight distribution across approaches compared to other methods.

5. Conclusions

We designed a range of attention-based architectures that could handle longitudinal data from EHRs and compared the discriminability and the clinical plausibility of the resulting attention maps. Each approach increased in complexity to accommodate more specific and informative attention mechanisms. The simplest model (time attention) could inform only the contribution of the time periods, serving as the basis for others that could also tell which clinical input features had the highest contribution (domain-specific attention), and finally, to one that was able to give the contribution of each feature at each time point (hierarchical attention).

The baseline model without attention performed the best in terms of discriminability, slightly decreasing the average performance from 0.98% to 8.7% as the architecture accommodated more informative attention mechanisms. Architectures from Approach 3 were the most informative as they could give the attention weights in two different domains (time and feature) simultaneously. Amongst the feature processing methods in this approach, the attention weights provided by the GRU were the most clinically plausible. In addition, this feature processing method achieved the most consistent results between approaches in terms of top-contributing features.

Some limitations should be considered. First, hyperparameter tuning was not performed and that could have modified discriminability. However, the same hyperparameters were used for all the combinations to achieve a fair comparison. Moreover, primary care EHRs contain a detailed and comprehensive picture of patients but lack information from other areas. Older patients might be reallocated to other healthcare facilities like nursing homes, partially losing their follow-up. Incorporating variables from other sources could have improved both discriminability and clinical plausibility. Finally, the evaluated architectures were complex and parameter-heavy and could have benefited from richer data distributions. Future work includes comparing these results with those obtained by other models like transformers and exploring the generalisability of these results to EHRs from different origins, other sizes of observation windows, and other prediction tasks. In addition, it should be noted that determining an "adequate overall performance" threshold from a single study may be difficult, as each healthcare application may need to optimize particular metrics over others. To define it, future work should consider performing this study in different use cases and consulting clinicians to establish criteria, perhaps through focus groups, on the determination of this threshold for acceptable discriminability, taking into account the loss or gain of interpretability of the model.

In conclusion, models handling longitudinal EHR data using simpler architectures achieved slightly better discriminability than those including more complex attention mechanisms for enhancing transparency. As Lipton noted, transparency may be at odds with predictive power [45] and thus the choice of the architecture should align with transparency requirements. In healthcare, prioritizing higher transparency and clinically plausible attention weights may justify choosing a model with a minor decrease in discriminability, provided overall performance remains adequate. Once considering including attention mechanisms, optimal combinations for domain-specific and hierarchical attention architectures were similar, utilizing either GRU or down-sampling CNN for feature processing. Hierarchical attention architectures, capable of highlighting clinical and temporal contributions, may be preferred for more informed decision-making when handling longitudinal EHR data.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/app15010146/s1, Figure S1: From the original source data to the input data matrix; Figure S2: Flow chart of the study population; Table S1: Description of the study population.

Author Contributions: M.C.-B., L.A.C.-R. and J.L.-J. conceptualised the study and defined the methodology. M.C.-B., L.A.C.-R. and C.V. participated in the funding acquisition. L.A.C.-R. performed the data curation, formal analysis, and visualisation. L.A.C.-R., M.C.-B. and J.L.-J. drafted the manuscript, and all authors participated in the review and editing of the final manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by a research grant from the Carlos III Institute of Health, Ministry of Economy and Competitiveness (Spain), awarded on the 2019 call under the Health Strategy Action 2013–2016, within the National Research Programme oriented to Societal Challenge and the Technical, Scientific and Innovation Research National Plan 2013–2016, (reference PI19/00535), and the PFIS Grant FI20/00040, co-funded with European Union ERDF (European Regional Development Fund) funds. This work has also been funded through project 6-SENSES (PID2022-138648OB-I00) funded by MCIN/AEI/10.13039/501100011033 and "ERDF A way of making Europe", EU. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Ethics Committee of Fundació Institut Universitari per a la recerca a l'Atenció Primària de Salut Jordi Gol i Gurina (IDIAPJGol) (#19/518-P).

Informed Consent Statement: The SIDIAP database is based on opt-out presumed consent. If a patient decides to opt-out, their routine data would be excluded from the database.

Data Availability Statement: Data sharing, distribution, or public availability is not allowed by European and national laws. However, researchers from public institutions can request data from SIDIAP. Further information is available online (https://www.sidiap.org/index.php/en/solicituds-en (accessed on 21 December 2024)). The code used to produce these results can be found on GitHub (www.github.com/IDIAPJGol/ComparingAttentionArchitectures (accessed on 21 December 2024)).

Acknowledgments: The authors thank Meggan Harris for her role in proofreading the manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CNN	Temporal Convolutional Neural Networks
DCNN	Down-sampling convolutional neural network
DUCNN	Down–up-sampling convolutional neural network
EHRs	Electronic Health Records
FC	Fully connected layer
GNN	Graph neural networks
GRU	Gated Recurrent Unit
RL	Recurrent layer
RNNs	Recurrent neural networks
ROC-AUC	Area under the ROC curve
PR-AUC	Area under the precision–recall curve

References

1. Cifuentes, M.; Davis, M.; Fernald, D.; Gunn, R.; Dickinson, P.; Cohen, D.J. Electronic Health Record Challenges, Workarounds, and Solutions Observed in Practices Integrating Behavioral Health and Primary Care. *J. Am. Board Fam. Med.* **2015**, *28*, S63–S72. [CrossRef] [PubMed]
2. Liu, F.; Panagiotakos, D. Real-world data: A brief review of the methods, applications, challenges and opportunities. *BMC Med. Res. Methodol.* **2022**, *22*, 287. [CrossRef] [PubMed]
3. Wang, S.; Gao, W.; Ngwa, J.; Allard, C.; Liu, C.T.; Cupples, L.A. Comparing baseline and longitudinal measures in association studies. *BMC Proc.* **2014**, *8*, S84. [CrossRef] [PubMed]
4. Nguyen, H.T.; Vasconcellos, H.D.; Keck, K.; Reis, J.P.; Lewis, C.E.; Sidney, S.; Lloyd-Jones, D.M.; Schreiner, P.J.; Guallar, E.; Wu, C.O.; et al. Multivariate longitudinal data for survival analysis of cardiovascular event prediction in young adults: Insights from a comparative explainable study. *BMC Med. Res. Methodol.* **2023**, *23*, 23. [CrossRef]
5. Cascarano, A.; Mur-Petit, J.; Hernández-González, J.; Camacho, M.; de Toro Eadie, N.; Gkontra, P.; Chadeau-Hyam, M.; Vitrià, J.; Lekadir, K. Machine and deep learning for longitudinal biomedical data: A review of methods and applications. *Artif. Intell. Rev.* **2023**, *56*, 1711–1771. [CrossRef]

6. Markus, A.F.; Kors, J.A.; Rijnbeek, P.R. The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. *J. Biomed. Inform.* **2021**, *113*, 103655. [CrossRef]
7. Arrieta, A.B.; Díaz-Rodríguez, N.; Ser, J.D.; Bennetot, A.; Tabik, S.; Barbado, A.; Garcia, S.; Gil-Lopez, S.; Molina, D.; Benjamins, R.; et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* **2020**, *58*, 82–115. [CrossRef]
8. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *arXiv* **2017**, arXiv:1706.03762.
9. Bahdanau, D.; Cho, K.; Bengio, Y. Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv* **2014**, arXiv:1409.0473.
10. Luong, M.T.; Pham, H.; Manning, C.D. Effective Approaches to Attention-based Neural Machine Translation. *arXiv* **2015**, arXiv:1508.04025.
11. Ross, A.S.; Hughes, M.C.; Doshi-Velez, F. Right for the Right Reasons: Training Differentiable Models by Constraining their Explanations. *arXiv* **2017**, arXiv:1703.03717.
12. Ribeiro, M.T.; Singh, S.; Guestrin, C. "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, 13–17 August 2016; KDD '16; pp. 1135–1144. [CrossRef]
13. Lundberg, S.M.; Lee, S.I. A Unified Approach to Interpreting Model Predictions. In *Proceedings of the Advances in Neural Information Processing Systems*; Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc.: San Francisco, CA, USA, 2017; Volume 30.
14. Zhang, J.; Kowsari, K.; Harrison, J.H.; Lobo, J.M.; Barnes, L.E. Patient2Vec: A Personalized Interpretable Deep Representation of the Longitudinal Electronic Health Record. *IEEE Access* **2018**, *6*, 65333–65346. [CrossRef]
15. Sha, Y.; Wang, M.D. Interpretable Predictions of Clinical Outcomes with An Attention-based Recurrent Neural Network. In Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, Boston, MA, USA, 20–23 August 2017; ACM: New York, NY, USA, 2017. [CrossRef]
16. Choi, E.; Bahadori, M.T.; Sun, J.; Kulas, J.; Schuetz, A.; Stewart, W. RETAIN: An Interpretable Predictive Model for Healthcare using Reverse Time Attention Mechanism. In Proceedings of the NIPS'16: Proceedings of the 30th International Conference on Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; Volume 29.
17. Nguyen, V.; Bodenreider, O. Adding an attention layer improves the performance of a neural network architecture for synonymy prediction in the UMLS Metathesaurus. *Stud. Health Technol. Inform.* **2022**, *290*, 116. [PubMed]
18. Carrasco-Ribelles, L.A.; Llanes-Jurado, J.; Gallego-Moll, C.; Cabrera-Bean, M.; Monteagudo-Zaragoza, M.; Violán, C.; del Olmo, E.Z. Prediction models using artificial intelligence and longitudinal data from electronic health records: A systematic methodological review. *J. Am. Med. Inform. Assoc.* **2023**, *30*, 2072–2082. [CrossRef]
19. Kabeshova, A.; Yu, Y.; Lukacs, B.; Bacry, E.; Gaïffas, S. ZiMM: A deep learning model for long term and blurry relapses with non-clinical claims data. *J. Biomed. Inform.* **2020**, *110*, 103531. [CrossRef]
20. Chen, P.; Dong, W.; Wang, J.; Lu, X.; Kaymak, U.; Huang, Z. Interpretable clinical prediction via attention-based neural network. *BMC Med. Inform. Decis. Mak.* **2020**, *20*, 1–9. [CrossRef]
21. Rodrigues-Jr, J.F.; Gutierrez, M.A.; Spadon, G.; Brandoli, B.; Amer-Yahia, S. LIG-Doctor: Efficient patient trajectory prediction using bidirectional minimal gated-recurrent networks. *Inf. Sci.* **2021**, *545*, 813–827. [CrossRef]
22. Rasmy, L.; Xiang, Y.; Xie, Z.; Tao, C.; Zhi, D. Med-BERT: Pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *Npj Digit. Med.* **2021**, *4*, 86. [CrossRef]
23. Collins, G.S.; Moons, K.G.M.; Dhiman, P.; Riley, R.D.; Beam, A.L.; Van Calster, B.; Ghassemi, M.; Liu, X.; Reitsma, J.B.; van Smeden, M.; et al. TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* **2024**, *385*, e078378. [CrossRef]
24. Baytas, I.M.; Xiao, C.; Zhang, X.; Wang, F.; Jain, A.K.; Zhou, J. Patient Subtyping via Time-Aware LSTM Networks. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, 13–17 August 2017; ACM: New York, NY, USA, 2017. [CrossRef]
25. Wang, T.; Qiu, R.G.; Yu, M. Predictive Modeling of the Progression of Alzheimer's Disease with Recurrent Neural Networks. *Sci. Rep.* **2018**, *8*, 9161. [CrossRef]
26. An, Y.; Zhang, L.; Yang, H.; Sun, L.; Jin, B.; Liu, C.; Yu, R.; Wei, X. Prediction of treatment medicines with dual adaptive sequential networks. *IEEE Trans. Knowl. Data Eng.* **2022**, *34*, 5496–5509. [CrossRef]
27. Jin, B.; Yang, H.; Sun, L.; Liu, C.; Qu, Y.; Tong, J. A Treatment Engine by Predicting Next-Period Prescriptions. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, London, UK, 19–23 August 2018; KDD '18; pp. 1608–1616. [CrossRef]

28. Catling, F.J.R.; Wolff, A.H. Temporal convolutional networks allow early prediction of events in critical care. *J. Am. Med. Inform. Assoc.* **2020**, *27*, 355–365. [CrossRef] [PubMed]
29. Li, Y.; Rao, S.; Solares, J.R.A.; Hassaine, A.; Ramakrishnan, R.; Canoy, D.; Zhu, Y.; Rahimi, K.; Salimi-Khorshidi, G. BEHRT: Transformer for Electronic Health Records. *Sci. Rep.* **2020**, *10*, 7155. [CrossRef] [PubMed]
30. Santana, A.; Colombini, E. Neural Attention Models in Deep Learning: Survey and Taxonomy. *arXiv* **2021**, arXiv:2112.05909.
31. Brauwers, G.; Frasincar, F. A General Survey on Attention Mechanisms in Deep Learning. *IEEE Trans. Knowl. Data Eng.* **2023**, *35*, 3279–3298. [CrossRef]
32. Xu, Q.; Duan, W. DualAttNet: Synergistic fusion of image-level and fine-grained disease attention for multi-label lesion detection in chest X-rays. *Comput. Biol. Med.* **2024**, *168*, 107742. [CrossRef]
33. Zhang, Z.; Gao, L.; Li, P.; Jin, G.; Wang, J. DAUF: A disease-related attentional UNet framework for progressive and stable mild cognitive impairment identification. *Comput. Biol. Med.* **2023**, *165*, 107401. [CrossRef]
34. Bibal, A.; Cardon, R.; Alfter, D.; Wilkens, R.; Wang, X.; François, T.; Watrin, P. Is Attention Explanation? An Introduction to the Debate. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Dublin, Ireland, 22–27 May 2022; Muresan, S., Nakov, P., Villavicencio, A., Eds.; pp. 3889–3900. [CrossRef]
35. Wiegreffe, S.; Pinter, Y. Attention is not not Explanation. *arXiv* **2019**, arXiv:1908.04626.
36. Jain, S.; Wallace, B.C. Attention is not Explanation. *arXiv* **2019**, arXiv:1902.10186.
37. Sen, C.; Hartvigsen, T.; Yin, B.; Kong, X.; Rundensteiner, E. Human Attention Maps for Text Classification: Do Humans and Neural Networks Focus on the Same Words? In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; Jurafsky, D., Chai, J., Schluter, N., Tetreault, J., Eds.; pp. 4596–4608. [CrossRef]
38. Riedl, M.O. Human-Centered Artificial Intelligence and Machine Learning. *arXiv* **2019**, arXiv:1901.11184. [CrossRef]
39. Zhou, P.; Shi, W.; Tian, J.; Qi, Z.; Li, B.; Hao, H.; Xu, B. Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Berlin, Germany, 7–12 August 2016; pp. 207–212. [CrossRef]
40. Recalde, M.; Rodríguez, C.; Burn, E.; Far, M.; García, D.; Carrere-Molina, J.; Benítez, M.; Moleras, A.; Pistillo, A.; Bolíbar, B.; et al. Data Resource Profile: The Information System for Research in Primary Care (SIDIAP). *Int. J. Epidemiol.* **2022**, *51*, e324–e336. [CrossRef] [PubMed]
41. GitHub: Comparing Discriminability and Attention Maps from Deep Learning Architectures Using Longitudinal Data from Electronic Health Records Repository. 2024. Available online: https://github.com/IDIAPJGol/ComparingAttentionArchitectures (accessed on 21 December 2024).
42. Calderón-Larrañaga, A.; Vetrano, D.L.; Onder, G.; Gimeno-Feliu, L.A.; Coscollar-Santaliestra, C.; Carfí, A.; Pisciotta, M.S.; Angleman, S.; Melis, R.J.; Santoni, G.; et al. Assessing and Measuring Chronic Multimorbidity in the Older Population: A Proposal for Its Operationalization. *J. Gerontol. Ser. A Biol. Sci. Med. Sci.* **2016**, *72*, 1417–1423. [CrossRef] [PubMed]
43. Orfila, F.; Carrasco-Ribelles, L.A.; Abellana, R.; Roso-Llorach, A.; Cegri, F.; Reyes, C.; Violán, C. Validation of an electronic frailty index with electronic health records: eFRAGICAP index. *BMC Geriatr.* **2022**, *22*, 404. [CrossRef] [PubMed]
44. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2017**, arXiv:1412.698.
45. Lipton, Z.C. The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* **2018**, *16*, 31–57. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.