Warfarin Dose Estimation on High-dimensional and Incomplete Data

Zeyuan Wang[1,2], Josiah Poon[1], Jie Yang[1], Simon Poon[1]*
[1]School of Computer Science, The University of Sydney, Sydney 2006, Australia
[2]AI Lab, Beijing Medicinovo Ltd., Beijing 100071, China
{zwan7221, jyan4704}@uni.sydney.edu.au, {josiah.poon, simon.poon}@sydney.edu.au

Abstract

Warfarin is a widely used oral anticoagulant worldwide. However, due to the complex relationship between individual factors, it is challenging to estimate the optimal warfarin dose to give full play to its ideal efficacy. Currently, there are plenty of studies using machine learning or deep learning techniques to help with the optimal warfarin dose selection. But few of them can resolve missing values and high-dimensional data naturally, that are two main concerns when analyzing clinical real world data. In this work, we propose to regard each patient's record as a set of observed individual factors, and represent them in an embedding space, that enables our method can learn from the incomplete date directly and avoid the negative impact from the high-dimensional feature set. Then, a novel neural network is proposed to combine the set of embedded vectors non-linearly, that are capable of capturing their correlations and locating the informative ones for prediction. After comparing with the baseline models on the open source data from International Warfarin Pharmacogenetics Consortium, the experimental results demonstrate that our proposed method outperform others by a significant margin. After further analyzing the model performance in different dosing subgroups, we can conclude that the proposed method has the high application value in clinical, especially for the patients in high-dose and medium-dose subgroups.

1. Introduction

Warfarin is a widely used anticoagulation for the treatment of non-valvular atrial fibrillation and venous thromboembolism [1]. However, due to its narrow therapeutic window and the large individual factor variability, especially for warfarin sensitive patients, it is difficult to deliver the optimal warfarin dose

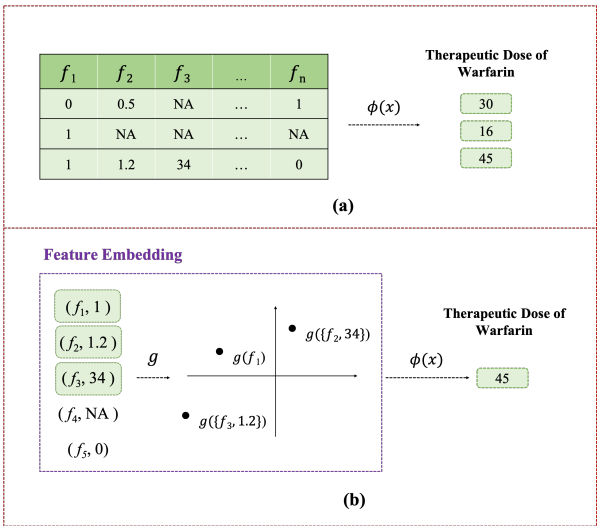

Figure 1. (a) An example of common data processing way. (b) An example of how we transforming and processing the original data from each individual patient.

[2], that the percentage of patients in the warfarin therapeutic window is even less than 60%, despite frequent use of INR for monitoring [3]. Therefore, the appropriate determination of warfarin dose is critical to its effectiveness and safety in clinical.

Till date, remarkable efforts have been invested to develop warfarin dose prediction models on the integration of clinical, demographic and genetic features for individual patients [1, 4–6], in which multivariate linear regression (MLR) is one of most common dosing algorithms [1, 7, 8]. However, linear models lack of effectiveness of learning non-linear relations and may not fit well to a certain subset of patients [9]. To avoid this, machine learning and deep learning approaches have been proposed recently, such as support vector machine (SVM) [10], decision tree based algorithms [11] and neural networks [4], that they are capable of capturing the complex relationships among individual factors and enhance the model performance. Both

* Corresponding Author

URI: https://hdl.handle.net/10125/71035
978-0-9981331-4-0

linear and machine learning models are constructed on the set of 1-dimensional vectors in some feature space (Shown in Figure 1a.), that always encounter two major challenges, high-dimension and missing values [12], because patients will not all examinations in hospital and their clinical features are various to each other, such as medications and indications. These two concerns are what our method aims to resolve.

For dealing with missing values, a typical strategy is to fill them by generating candidates from the existing data distribution, such as the maximum value, average value, and candidates computed by MICE or KNN [13–15]. When the missing rate remains at the low level, data imputation methods are with high accuracy. However, they are not solid on the highly incomplete data sets, such as more than 50% of the features with more than 80% missing rate [16], since the observed data is not able to represent the overall distribution and imputation techniques will bring much extra noise to interfere with the final prediction. Instead of filling missing values, predictive models can be constructed under certain assumptions of missing mechanisms without imputation required in prior, and their decision functions can be relied on the witnessed data only [17–19]. While, all these methods are based on the fixed feature space, which will always be high-dimensional, especially in the data sets from real-world settings, and their performance will be negatively influenced. In our study, we not only learn from the incomplete data directly, but also resolve the high-dimension problem.

The common way for processing high-dimensional data set is feature selection to enhance the predictive performance, provide more cost-effective models, and provide the understanding of underlying data pattern [20]. There are three main strategies of feature selection [21], filter methods, wrapper methods, and embedded methods. The key concept in filter methods is feature ranking or ordering by the feature relevance measurement [22–24], and different from it, wrapper methods use predictive performance as the objective function to evaluate the usefulness of feature subset [25, 26]. As for embedded methods, they are developed to reduce the computation time when reclassifying different subsets from wrapper methods, and incorporate feature selection in training process [27, 28].

In this work, we address both high-dimensional feature space problem and missing values at the same time in an alternative way, that we view each patient's record as a set of observed features and map them to an embedding space via feature embedding (Shown in Figure 1b.). Through this way, our proposed method can learn from incomplete and high-dimensional data directly and naturally without any initial operation. After incorporating with a novel neural network, which is capable of exploring the correlations among the embedded vectors from the observed features and robust to the large amount of invalid information, the containing parameters are trained jointly in an end-to-end manner.

2. Related Work

2.1. Transformer

In this work, in order to capture the complex relationships of the embedded vectors from observed features, we use a state-of-art technique from the NLP community, *Transformer* [29], that is originally proposed for the neural machine translation (NMT) tasks and use self-attention mechanism to resolve the long distance dependency problem. Currently, it has been widely applied in real-world applications such as recommender system [30], automatic knowledge graph construction [31] and speech recognition [32, 33].

The key idea of *Transformer* is the proposed multi-head attention, that enables the model to capture associations between the input words in different embedding subspaces, named as heads. Moreover, there exists a novel computational module, feed-forward network (FFN), *in Transformer* to further enhance the representation abilities of the vectors from multi-head attention module [34].

In our study, we consider each patient's record as a set of observed features. Through feature embedding, multi-head attention and FFN in *Transformer*, an optimal feature representation space for warfarin dose prediction is able to obtained. Notably, when we use techniques from *Transformer*, the position encoding vectors of words are excluded, since there is no sequential information existing.

2.2. Multi-Instance Pooling

As the key step in multi-instance neural network (MINN) [35], multi-instance pooling is to obtain the informative bag or instance representation, that bridges the bag space and instance space [36]. Typically, there are two main strategies of adopting the multi-instance pooling on MINN, that are trainable ones and non-trainable ones [37].

Non-trainable multi-instance pooling methods are the most common in MINN, such as max pooling [35], average pooling [38] and sum pooling [39]. They use straightforward data operation to obtain their appropriate representations and uncover the hidden patterns among instances and bags. In addition, there are some novel trainable multi-instance pooling

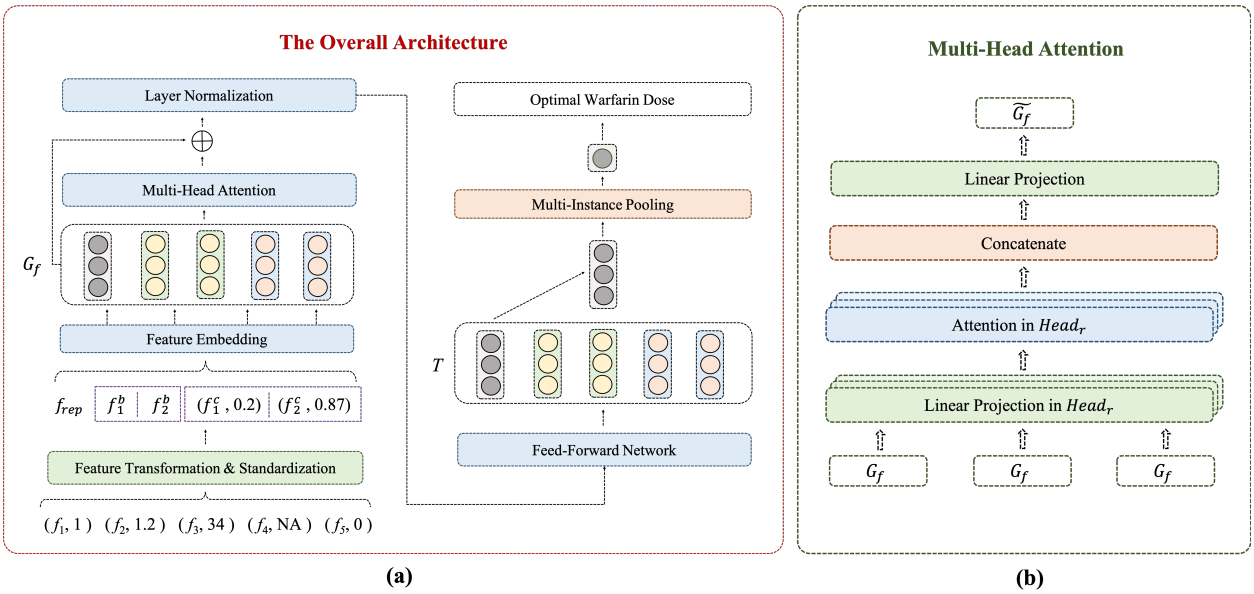

Figure 2. (a) The overall architecture of our proposed method, that consists of two levels, a feature embedding module to map the observed information to en embedding space and a novel neural network with multi-head attention, feed-forward network and multi-instance pooling to fully explore the hidden patterns among features for the final prediction. The whole process is trained jointly in end-to-end. (b) Multi-head attention is adopted in our method to uncover the correlations of embedded vectors, i.e., observed features. Notably, through multi-head attention the relations between G_{rep} and the other embedded vectors can also be captured, which is used to represent the overall patient information for the further processing.

methods proposed, such as gated attention based MIL pooling [40], attention based MIL pooling [40] and dynamic pooling [36]. These pooling methods can help MINNs with the key instance selection through assigning different weights for them. Multi-instance pooling techniques, no matter trainable or non-trainable ones, enable the model to avoid the negative influence introduced by the invalid information to enhance the performance.

Normally, MINN is a typical framework for classification tasks and multi-instance pooling plays a crucial role, while in our work, we adopt it for regression, that we regard the final representation vector as a bag of instances and use max pooling to compute the bag representation for the final warfarin dosing prediction.

3. Methodology

In order to estimate the appropriate warfarin dose on incomplete and high-dimensional data, we propose a novel framework, that each patient $(X,\ y)$ is initially transformed to a set of observed feature-value pairs $X = \{(f_1, v_1), (f_2, v_2), \ldots, (f_n, v_n)\}$ with the corresponding optimal warfarin dose y, and feature f_j $(j = 1, 2, \ldots, n)$ is either binary f_j^b, ordinal or continuous f_j^c. Notably, all nominal features are one-hot encoded to binary ones during transformation process. Our objective is to train a regressor to predict y from the feature set X. Specifically, our modeling strategy consists of two levels, that the underlying part is to represent each observed feature-pair (f_j, v_j) by a d dimensional embedded vector $g_f = g(f_j, v_j) \in \mathbb{R}^d$. The second level is a novel neural network to capture the complex correlations in G and locate the valuable information in G for the warfarin optimal dose $y \in \mathbb{R}$ estimation, where $G = \{g_{rep}, g(f_1, v_1), g(f_2, v_2), \ldots, g(f_n, v_n)\}$ and g_{Rep} is an embedded vector of a representation feature, which is added in the feature set representing all observed information, i.e., the overall body condition. These two parts are parameterized and trained jointly in an end-to-end manner. The overall architecture is shown in Figure 2a.

3.1. Feature Embedding

For each patient, the observed features is first transformed to the feature-value pairs and in order to make them applicable to feature embedding, we design

two different strategies in terms of their different types:

- **Binary and Nominal Features:** As mentioned above, nominal features are initially one-hot encoded to binary ones and in feature embedding, we only include the binary features f^b with the positive responses, i.e., only the exposure factors are included into predictive modeling.
- **Continuous and Ordinal Features:** As for continuous and ordinal features, we standardize and scale their corresponding vaues to 0 to 1 by $\left(f^c(i) - \min\left(f^c\right)\right) / \left(\max\left(f^c\right) - \min\left(f^c\right)\right)$ where $f^c(i)$ is the i_{th} sample in the continuous or ordinal feature f^c.

After preprocessing, each patient X can be denoted as $X = \{f^b_{rep}, f^b_1, \ldots, f^b_{n_1}, (f^c_1, v_1), \ldots, (f^c_{n_2}, v_{n_2})\}$, where f^b_{rep} is a self-defined binary feature with positive response. Through embedding and multi-head attention, f^b_{rep} can represent the overall body condition and be used for the final prediction. Subsequently, f^b and f^c are embedded via two different ways, that for f^b, we parameterize it through:

$$g(f^b) = L_{f^b} \text{ where } L_{f^b} \in \mathbb{R}^d \tag{1}$$

L_{f^b} is a d-dimensional parameter vector, which optimized by the back propagation. As for f^c, we map it to the embedding space by:

$$g(f^c, v) = W^c(vL_{f^c}/d) \tag{2}$$

where $L_{f^c} \in \mathbb{R}^d$ is also a parameter vector of d dimensions and v is the corresponding value of f^c. $W^c \in \mathbb{R}^{d \times d}$ is used as capacity control to adjust the contributions of the value v on the parameter vector L_{f^c}.

At last, we combine all obtained embedded vectors from f^b and f^c respectively, and adopt the layer normalization [41] as our feature embedding output:

$$G_f = \text{LayerNorm}\left(\left[g(f^b), g(f^c, v)\right]\right) \tag{3}$$

Through feature embedding, we can map all observed feature to an informative embedding space, that the negative effects of missing values and high dimensional features are also naturally avoided, and leaves great flexibility to analyze the potential associations between features by the following technique, multi-head attention.

3.2. Multi-Head Attention

The multi-head attention module we used is identical to the one originally defined by [29] except that we exclude the position encoding part, since there is no sequential information existing. There are two computation part in multi-head attention, scaled dot-product attention and multi-head transformation. In scaled dot-product attention, three input vectors with d_k dimensions are received, a query Q, key K and the corresponding value V and the output is obtained by:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{4}$$

In this work, we mainly focus on the mining of potential associations between observed features, so Q, K, V are all the embedded vector G_f from the previous feature embedding:

$$\text{Attention}(G_f, G_f, G_f) = \text{softmax}\left(\frac{G_f G_f^T}{\sqrt{d_k}}\right)G_f \tag{5}$$

In addition, to fully uncover the underlying relations of G_f, multi-head attention allows to access several sub-embedding spaces via multi-head transformation:

$$H_r = \text{Attention}\left(W_r^1 G_f, W_r^2 G_f, W_3^1 G_f\right) \tag{6}$$

where W^r is the output of a single attention head and W_r^1, W_r^2, W_r^3 are three linear projections for G_f. Then they are concatenate as the final output of multi-head attention:

$$\text{MultiHead}(G_f, G_f, G_f) = [H_1; \ldots; H_R]\, W^4 \tag{7}$$

where W^4 is the output projection. Through this way, the associations between the self-defined representation vector g_{rep} and other observed features can be explored, that allows g_rep can represent the patient's current medical condition and be used for the further warfarin dose estimation.

3.3. Feed-Forward Network

After multi-head attention, we adopt feed-forward network, consisting of two 1-dimensional convolution layers with kernel size equals to 1 and ReLu activation in between, to further enhance the representation capability of G_f:

$$\text{FFN}(x) = \text{Conv1D}(\max\left(0, \text{Conv1D}(x)\right)) \tag{8}$$

Through feed-forward network we can obtain our final vector set $T = \{t_{rep}, t_1, t_2, \ldots, t_n\}$ and we only include the self-defined representation vector t_{rep} with d dimensions for the further operation, multi-instance pooling.

3.4. Multi-Instance Pooling

As the key step in MINN, multi-instance pooling enables models to avoid the interference from noise and invalid information [39]. Out of this consideration, we adopt multi-instance pooling instead of common used fully connected layer as our final output layer to get the predicted warfarin dose. Moreover, our objective is a regression task and the trainable multi-instance pooling methods typically generates weights between 0 and 1, that are more suitable for classification instead of regression tasks, so we adopt a non-trainable method, max pooling.

$$\text{Output} = \max\left(t_k = \{w_1, w_2, \ldots, w_d\}\right) \qquad (9)$$

$w_i \in \mathbb{R}$ is a parameter in the representation vector t_k.

After we obtain the final output, we compute the loss with the true value y and train our model by back propagation in end-to-end way.

4. Experiments

4.1. Data Description

The data we used for modeling is the open source IWPC cohort which has been described previously by [1] and can be downloaded from the PharmGKB website (http://www.pharmgkb.org/downloads/). The data set contains 6256 warfarin users from 4 continents with demographic factors, clinical features, such as age, weight, height, indications and united medication, as well as CYP2C9 and VKORC1 genotypes. We exclude the subjects without reaching stable doses of warfarin and the therapeutic doses of warfarin are missing. A total of 5410 subjects are included in our study. Moreover, we include all 8 indications, 1458 comorbidities and 1917 medications, so the dimensions of the dataset are in high level. Besides, the data set is with a large number of missing values with 41.8% missing rate on average. The detailed statistics of our included data are shown in Table 1.

4.2. Experimental Settings

We map feature-pairs to an embedding space with 512 dimensions and processed by the multi-head attention with 8 heads, i.e., 8 embedding sub-spaces to mine underlying relationships. In feed-forward network, we set two convolution layers with 1024 and 512 dimensions, respectively. Moreover, in each computational module, the dropout layer [42] is added with the 0.3 dropout rate to prevent model overfitting. Our method is trained by Adam optimizer [43] with 300 epochs and $5e^{-6}$, that ϵ is $1e^{-8}$ and the momentum parameters β_1, β_2 are set to 0.9 and 0.98. Moreover, for the sake of fairly comparison, we design "early stopping" mechanism on five-fold cross validation in terms of R_2, MAE (Mean Absolute Error) and MSE (Mean Squared Error). The loss function we used is Log-Cosh loss, which is defined as :

Table 1. Statistics of the Data Set

Included Patients	5410
Binary or Nominal Features	3395
Continuous or Ordinal Features	4
Max. Observed Features	58
Min. Observed Features	5
Average Missing Rate	41.8%
Max. Missing Rate	83.8%

$$loss(y, f(x)) = \sum_{i=1}^{n} \log\cosh(y_{true} - y_{pred}) \qquad (10)$$

where y_{true} is the true value and y_{pred} denotes the predicted value on the i^{th} sample.

4.3. Baselines

To comprehensively evaluate our proposed method, we first compare it with three advanced machine learning methods XGBoost [44], LightGBM [45] and CatBoost [46], and to obtain the best performance of machine learning performance, we use AuoML method [47] to select the best parameter set automatically. They are all decision tree based methods, not only can learn from incomplete data directly, but also robust to sparse data, that are exactly two challenges we meet.

In addition, to demonstrate the effectiveness of what we used multi-instance pooling method, we conduct the performance comparisons by using fully connected layer, max pooling, mean pooling [39], attention based pooling, and gated attention based pooling methods [40] in our proposed framework.

5. Results and Analysis

5.1. Performance Comparisons

We first compare our proposed method with three advanced machine learning techniques, and then we evaluate the performance of our method with different pooling methods to generate the final output. The detailed comparison results are shown in Table2, that we

Table 2. Comparison with Baseline Methods for Warfarin Dose Estimation

Strategy	Models	R^2	MAE	MSE
Machine Learning Techniques	XGBoost-AutoML	0.420	8.979	167.972
	LightGBM-AutoML	0.327	9.495	190.347
	CatBoost-AutoML	0.427	8.773	163.884
With Different Pooling Methods	Fully Connected Layer	0.405	8.742	168.549
	Mean Pooling	0.418	8.626	165.069
	Att. Pooling	0.426	8.639	163.122
	Gated Att. Pooling	0.424	8.580	163.558
This Work	Max Pooling	**0.437**	**8.471**	**160.016**

use R^2, MAE and MSE as our evaluation metrics. The larger the R^2, the better, MAE and MSE are opposite.

As demonstrated, our proposed method achieves consistent better results compared with all baseline methods, that the R^2, MAE and MSE are 0.437, 8.471 and 160.016, respectively. Surprisingly, CatBoost performs much better than XGBoost and LightGBM, illustrating its superiority of processing categorical and sparse matrix, which is one of the characteristics of the data we use. Moreover, the comparison results of fully connected layer and other multi-instance pooling methods, shows the deficiencies of fully connected layer in effective information location, that its R^2 is only 0.405. As for attention based (Att. Pooling) and gated attention based multi-instance pooling (Gated Att. Pooling), the key idea is to assign different weights for instances to adjust their contributions, but the weights are all from 0 to 1, which may limit the model performance on regression tasks. The R^2 for Att. Pooling and Gated Att. Pooling are 0.426 and 0.424, respectively, lower than our used max pooling.

5.2. Impact of Multi-Head Attention

To measure the impact that the multi-head attention introduces, we conduct experiments for our method with 0, 2, 4, 6, 8, 10, and 12 heads, respectively, where 0 heads denotes that multi-head attention is not employed and integrated in our proposed neural network. The evaluation results are depicted in Figure 3.

The number of 8 heads gives the best model performance, i.e., relations capture in 8 subspaces can be fully uncovered, which brings clinical instructions that the features we used can be considered in 8 aspects and in each one, features are with their own hidden associations. More importantly, our experiments prove the effectiveness of multi-head attention in our method

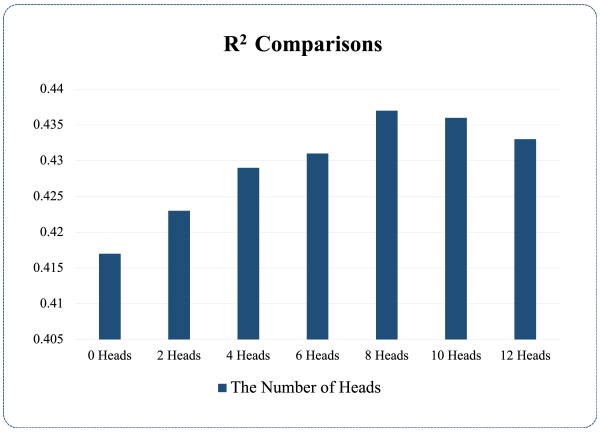

Figure 3. The Performance Comparisons of Different Number of Heads

since if we remove it, the R^2 drop to 0.417. It also illustrates the complex relations between clinical features existing in warfarin users, and correlations exploration is necessary.

5.3. Dose Subgroup Analysis

In this subsection, we mainly focus on evaluating the clinical applicability of our method on different dose subgroups. Follow the same criteria described in [1], warfarin doses are divided into low dose group $\leq 21\text{mg/wk}$, medium dose group > 21 to $< 49\text{mg/wk}$, and high dose group $\geq 49\text{mg/wk}$. In different subgroups, we measure the model clinical applicability using ideal predictive percentage (IPP) [1], indicating the percentages of the predicted dose within the 20% interval of the actual dose.

As shown in Figure 4, our method is with high clinical applicability, especially in the medium and high

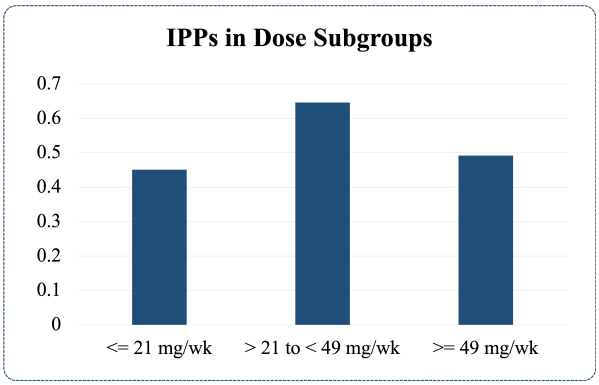

Figure 4. The IPPs Comparison in Different Dose Subgroups

Table 3. Comparisons of Our Method on Different Feature Sets

Feature Set	R^2	MAE	MSE
CF	0.412	8.612	163.559
Cl	0.383	8.972	174.648
GF	0.395	8.910	169.513
This Work (Mixed)	**0.437**	**8.471**	**160.016**

dose group, the IPPs are 0.646 and 0.492, respectively. Moreover, high and low dose groups are with lower IPP than the medium group, revealing that patients in medium dose group are with less clinical variability and more stable disease condition, making them easier to obtain the optimal warfarin dose from the predictive models.

5.4. Different Feature Sets

At last, we evaluate our method on four different combinations of features, only continuous or ordinal features (CF), only clinical factors with the exclusion of CF and genetic variables (Cl), only the genetic features (GF), and all features (Mixed). Table 3 records the comparison results detaily.

As shown, compared to the other clinical features, the continuous or ordinal features, such as age, weight, height, target INR, and genetic variables provides the main guidance and instruction for optimal warfarin dose prediction. However, clinical features also contain fruitful information for dose estimation, so mixing them together to obtain the comprehensive clinical guidance is the best option for modeling.

6. Conclusion

This paper presents a novel and applicable way for warfarin dose estimation on incomplete and high-dimensional data, that data imputation and feature selection is not required priorly through transforming observed information into feature-value pairs and modeling on them. Specifically, it consists of two levels, and the first level is a feature embedding module to map all observed information to an embedding space to avoid missing values and redundant features naturally. Based on the embedded vectors, the second modeling level is developed by a novel neural network, that not only can capture the underlying and complex relationships among features, but also can isolate the invalid information and noise to make the better warfarin dose determination.

The two main contributions of our method are feature embedding and multi-instance pooling. Through feature embedding, we can obtain an informative embedding space, that leaves great flexibility of further operation. For example, we can use many NLP techniques to discover our interested information, and the only thing to notice is the removal of the sequential information. The other one is the multi-instance pooling, that it is capable of protect our model from large amount of invalid information. More importantly, multi-instance pooling is normally used for the classification tasks, but in our method, we have demonstrated its feasibility of applying to regression problems, that the application of this technique has been expanded.

In the future work, we try to expand our current method to deal with temporal data, such as frequent physical test results and lab test results, and discover more hidden patterns among them.

References

[1] I. W. P. Consortium, "Estimation of the warfarin dose with clinical and pharmacogenetic data," *New England Journal of Medicine*, vol. 360, no. 8, pp. 753–764, 2009.

[2] Y.-b. Zhu, X.-h. Hong, M. Wei, J. Hu, X. Chen, S.-k. Wang, J.-r. Zhu, F. Yu, and J.-g. Sun, "Development of a novel individualized warfarin dose algorithm based on a population pharmacokinetic model with improved prediction accuracy for chinese patients after heart valve replacement," *Acta Pharmacologica Sinica*, vol. 38, no. 3, pp. 434–442, 2017.

[3] W.-y. Shu, J.-l. Li, X.-d. Wang, and M. Huang, "Pharmacogenomics and personalized medicine: a review focused on their application in the chinese population," *Acta pharmacologica Sinica*, vol. 36, no. 5, pp. 535–543, 2015.

[4] E. Grossi, G. M. Podda, M. Pugliano, S. Gabba, A. Verri, G. Carpani, M. Buscema, G. Casazza, and M. Cattaneo, "Prediction of optimal warfarin

maintenance dose using advanced artificial neural networks," *Pharmacogenomics*, vol. 15, no. 1, pp. 29–37, 2014.

[5] P. Lenzini, M. Wadelius, S. Kimmel, J. Anderson, A. Jorgensen, M. Pirmohamed, M. Caldwell, N. Limdi, J. Burmester, M. Dowd, *et al.*, "Integration of genetic, clinical, and inr data to refine warfarin dosing," *Clinical Pharmacology & Therapeutics*, vol. 87, no. 5, pp. 572–578, 2010.

[6] R. Liu, X. Li, W. Zhang, and H.-H. Zhou, "Comparison of nine statistical model based warfarin pharmacogenetic dosing algorithms using the racially diverse international warfarin pharmacogenetic consortium cohort database," *PloS one*, vol. 10, no. 8, p. e0135784, 2015.

[7] T. Gaikwad, K. Ghosh, P. Avery, F. Kamali, and S. Shetty, "Warfarin dose model for the prediction of stable maintenance dose in indian patients," *Clinical and Applied Thrombosis/Hemostasis*, vol. 24, no. 2, pp. 353–359, 2018.

[8] B. F. Gage, C. Eby, P. E. Milligan, G. A. Banet, J. R. Duncan, and H. L. McLeod, "Use of pharmacogenetics and clinical factors to predict the maintenance dose of warfarin," *Thrombosis and haemostasis*, vol. 91, no. 01, pp. 87–94, 2004.

[9] Z. Ma, P. Wang, Z. Gao, R. Wang, and K. Khalighi, "Ensemble of machine learning algorithms using the stacked generalization approach to estimate the warfarin dose," *PloS one*, vol. 13, no. 10, p. e0205872, 2018.

[10] E. Cosgun, N. A. Limdi, and C. W. Duarte, "High-dimensional pharmacogenetic prediction of a continuous trait using machine learning techniques with application to warfarin dose prediction in african americans," *Bioinformatics*, vol. 27, no. 10, pp. 1384–1389, 2011.

[11] K. Liu, C.-L. Lo, and Y.-H. Hu, "Improvement of adequate use of warfarin for the elderly using decision tree-based approaches," *Methods of Information in Medicine*, vol. 53, no. 01, pp. 47–53, 2014.

[12] J. Lin, T. Jiao, J. E. Biskupiak, and C. McAdam-Marx, "Application of electronic medical record data for health outcomes research: a review of recent literature," *Expert review of pharmacoeconomics & outcomes research*, vol. 13, no. 2, pp. 191–200, 2013.

[13] A. R. T. Donders, G. J. Van Der Heijden, T. Stijnen, and K. G. Moons, "A gentle introduction to imputation of missing values," *Journal of clinical epidemiology*, vol. 59, no. 10, pp. 1087–1091, 2006.

[14] G. E. Batista, M. C. Monard, *et al.*, "A study of k-nearest neighbour as an imputation method.," *His*, vol. 87, no. 251-260, p. 48, 2002.

[15] S. v. Buuren and K. Groothuis-Oudshoorn, "mice: Multivariate imputation by chained equations in r," *Journal of statistical software*, pp. 1–68, 2010.

[16] W. R. Hersh, M. G. Weiner, P. J. Embi, J. R. Logan, P. R. Payne, E. V. Bernstam, H. P. Lehmann, G. Hripcsak, T. H. Hartzog, J. J. Cimino, *et al.*, "Caveats for the use of operational electronic health record data in comparative effectiveness research," *Medical care*, vol. 51, no. 8 0 3, p. S30, 2013.

[17] Z. Ghahramani and M. I. Jordan, "Supervised learning from incomplete data via an em approach," in *Advances in neural information processing systems*, pp. 120–127, 1994.

[18] P. K. Shivaswamy, C. Bhattacharyya, and A. J. Smola, "Second order cone programming approaches for handling missing and uncertain data," *Journal of Machine Learning Research*, vol. 7, no. Jul, pp. 1283–1314, 2006.

[19] X. Liao, H. Li, and L. Carin, "Quadratically gated mixture of experts for incomplete data classification," in *Proceedings of the 24th International Conference on Machine learning*, pp. 553–560, 2007.

[20] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of machine learning research*, vol. 3, no. Mar, pp. 1157–1182, 2003.

[21] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Computers & Electrical Engineering*, vol. 40, no. 1, pp. 16–28, 2014.

[22] P. Langley *et al.*, "Selection of relevant features in machine learning," in *Proceedings of the AAAI Fall symposium on relevance*, vol. 184, pp. 245–271, 1994.

[23] R. Battiti, "Using mutual information for selecting features in supervised neural net learning," *IEEE Transactions on neural networks*, vol. 5, no. 4, pp. 537–550, 1994.

[24] H. Stoppiglia, G. Dreyfus, R. Dubois, and Y. Oussar, "Ranking a random feature for variable and feature selection," *Journal of machine learning research*, vol. 3, no. Mar, pp. 1399–1414, 2003.

[25] P. M. Narendra and K. Fukunaga, "A branch and bound algorithm for feature subset selection," *IEEE Transactions on computers*, no. 9, pp. 917–922, 1977.

[26] J. Reunanen, "Overfitting in making comparisons between variable selection methods," *Journal of Machine Learning Research*, vol. 3, no. Mar, pp. 1371–1382, 2003.

[27] O. Chapelle and S. S. Keerthi, "Multi-class feature selection with support vector machines," in *Proceedings of the American statistical association*, vol. 58, 2008.

[28] J. Neumann, C. Schnörr, and G. Steidl, "Combined svm-based feature selection and classification," *Machine learning*, vol. 61, no. 1-3, pp. 129–150, 2005.

[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, pp. 5998–6008, 2017.

[30] S. Zhang, L. Yao, A. Sun, and Y. Tay, "Deep learning based recommender system: A survey and new perspectives," *ACM Computing Surveys (CSUR)*, vol. 52, no. 1, pp. 1–38, 2019.

[31] A. Bosselut, H. Rashkin, M. Sap, C. Malaviya, A. Celikyilmaz, and Y. Choi, "Comet: Commonsense transformers for automatic knowledge graph construction," *arXiv preprint arXiv:1906.05317*, 2019.

[32] C.-C. Chiu, T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, E. Gonina, *et al.*, "State-of-the-art speech recognition with sequence-to-sequence models," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4774–4778, IEEE, 2018.

[33] T. Hayashi, S. Watanabe, T. Toda, and K. Takeda, "Multi-head decoder for end-to-end speech recognition," *arXiv preprint arXiv:1804.08050*, 2018.

[34] D. W. Otter, J. R. Medina, and J. K. Kalita, "A survey of the usages of deep learning for natural language processing," *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

[35] X. Wang, Y. Yan, P. Tang, X. Bai, and W. Liu, "Revisiting multiple instance neural networks," *Pattern Recognition*, vol. 74, pp. 15–24, 2018.

[36] Y. Yan, X. Wang, X. Guo, J. Fang, W. Liu, and J. Huang, "Deep multi-instance learning with dynamic pooling," in *Asian Conference on Machine Learning*, pp. 662–677, 2018.

[37] Z. Wang, J. Poon, and S. Poon, "Ami-net+: A novel multi-instance neural network for medical diagnosis from incomplete and imbalanced data," *arXiv preprint arXiv:1907.01734*, 2019.

[38] Y. Wang, J. Li, and F. Metze, "A comparison of five multiple instance learning pooling functions for sound event detection with weak labeling," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 31–35, IEEE, 2019.

[39] Z. Wang, J. Poon, S. Sun, and S. Poon, "Attention-based multi-instance neural network for medical diagnosis from incomplete and low quality data," in *2019 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, IEEE, 2019.

[40] M. Ilse, J. M. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," *arXiv preprint arXiv:1802.04712*, 2018.

[41] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[42] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[43] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[44] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

[45] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," in *Advances in neural information processing systems*, pp. 3146–3154, 2017.

[46] A. V. Dorogush, V. Ershov, and A. Gulin, "Catboost: gradient boosting with categorical features support," *arXiv preprint arXiv:1810.11363*, 2018.

[47] C. Thornton, F. Hutter, H. H. Hoos, and K. Leyton-Brown, "Auto-weka: Combined selection and hyperparameter optimization of classification algorithms," in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 847–855, 2013.