An Interpretable Deep Embedding Model for Few and Imbalanced Biomedical Data

Haishuai Wang, Li Li, Jiali Ma, Lianhua Chi, Jianjun Yang, Guangyu Tao, Jun Wu, Ziping Zhao, George M. Church

***Abstract*—In healthcare, training examples are usually hard to obtain (e.g., cases of a rare disease), or the cost of labelling data is high. With a large number of features (p) be measured in a relatively small number of samples (N), the "big p, small N" problem is an important subject in healthcare studies, especially on the genomic data. Another major challenge of effectively analyzing medical data is the skewed class distribution caused by the imbalance between different class labels. In addition, feature importance and interpretability play a crucial role in the success of solving medical problems. Therefore, in this paper, we present an interpretable deep embedding model (IDEM) to classify new data having seen only a few training examples with highly skewed class distribution. IDEM model consists of a feature attention layer to learn the informative features, a feature embedding layer to directly deal with both numerical and categorical features, a siamese network with contrastive loss to compare the similarity between learned embeddings of two input samples. Experiments on both synthetic data and real-world medical data demonstrate that our IDEM model has better generalization power than conventional approaches with few and imbalanced training medical samples, and it is able to identify which features contribute to the classifier in distinguishing case and control.**

***Index Terms*—Interpretable AI, Deep Embedding Model, Few Medical Data, Imbalanced Medical Data, Siamese Network**

I. Introduction

In the era of big data, data is involving all aspects of human life, including biology and medicine. Deep learning techniques provide effective paradigms to build end-to-end systems from complex and high-dimensional biomedical data, including electronic health records, omics and medical imaging [1]. However, differ from humans that can recognize new object classes from very few instances, most deep learning techniques require thousands of samples to achieve good performance. Thus, deep learning succeeds in data-intensive applications, but it lacks the ability of learning from a limited number of samples. Since the cost of labelling medical data is high [2], there are usually only few samples are available in healthcare [3]. Apart from few samples issue, the registered medical data are often highly imbalanced because cases of a rare disease are usually hard to obtain, resulting in highly skewed class distribution problem. Conventional classifiers typically perform poorly in imbalanced data, as they implicitly give the same attention to the majority class and the minority class. However, for medical related data analytics, accurately detecting minority class is of great importance since they correspond to high-impact events [4].

In addition, feature importance and interpretability of deep learning models play a crucial role in the success of solving real problems in healthcare. Medical datasets usually consist of a large number of disease markers, while some disease markers are not helpful and sometimes even have negative effects for clinical analysis. Nevertheless, most deep learning models are learned as a black box due to its high inherent complexity. Although deep learning demonstrates high performance and time-efficiency, researchers are not sure exactly which features contribute to the model used to classify diagnoses. The way those deep learning algorithms arrive at their conclusions needs to be understandable and interpretable. Moreover, deep learning algorithms tend to overweigh peripheral features at the expense of critical ones when they try to take all factors into account, resulting in inapplicability or overfitting problems on high-dimensional genomic data with only few samples. Therefore, integrating feature selection is necessary as it enables to remove those unimportant disease markers and select contributory ones.

Motivated by the above observations, our research mainly aims to solve the following challenges:

- How to build an interpretable deep learning model that can explain its decision-making to physicians by identifying relevant disease markers and facilitating clinical translation, is one of the challenges.
- Medical data are usually a mix of numerical and categorical features, which pose a challenge for directly applying classifiers as they can only handle numerical inputs by design. How to learn deep embedding representations from heterogeneous medical data is another challenge.
- Deep learning succeeds in data-intensive applications, but it lacks the ability of learning from a limited number of samples or imbalanced data. How to train a deep learning model using few samples on the imbalanced data is also challenging.

H. Wang is with Zhejiang Provincial Key Laboratory of Service Robot, College of Computer Science, Zhejiang University, Hangzhou, China.

L. Li and G. Church are with the Department of Genetics, Blavatnik Institute, and Wyss Institute for Biologically Inspired Engineering, Harvard University, Boston, MA, USA.

L. Chi is with the Department of Computer Science, La Trobe University, Melbourn, VIC, Australia.

J. Ma and Z. Zhao are with College of Computer Science, Tianjin Normal University, Tianjin, China.

J. Yang is with Department of General Practice, Shandong Provincial Third Hospital, Cheeloo College of Medicine, Shandong University, Jinan, China.

G. Tao is with the Department of Radiology, Shanghai Chest Hospital, Shanghai Jiao Tong University, Shanghai, China.

J. Wu is with School of Computer Science, Beijing Jiaotong University, Beijing China.

Co-first authors: Haishuai Wang and Li Li.

Corresponding authors: Haishuai Wang and Jianjun Yang.

To tackle the above challenges, in this paper, we propose an interpretable deep embedding model (IDEM) for few and imbalanced medical data analysis. The main contributions of our work are summarised as follows:

- A new interpretable deep embedding model is devised that integrates feature attention, categorical feature embedding and trains on few samples and imbalanced data. The proposed model is able to learn which features are informative in a deep neural network architecture.
- We formulate the classification problem as a verification task. IDEM learns deep embeddings of inputs, and then an identity network is used and a distance function is learned between their embeddings to output their similarity. This is done by applying cosine distance on the output embeddings and adding one fully connected layer to learn the weighted distance.
- The proposed model is able to explain its decision-making to physicians by identifying relevant clinical features and showing how much do these features contribute on diagnosis, facilitating clinical translation. By providing feedback on the importance of various clinical features in performing differential diagnosis, our model have the potential to improve clinical practicality.

We also evaluate our proposed algorithm on one synthetic data and three real medical datasets, and compare it with the state-of-the-art approaches for few and imbalanced medical data analysis. The experimental results demonstrate the effectiveness of the proposed model.

The remainder of the paper is organized as follows. Section II discusses related work. Section III presents the proposed IDEM model. Section IV reports experimental results. We conclude the paper in Section V.

II. Related work

In this section, we review existing research related to our work in the following areas: feature selection, feature embedding, learning from few samples and imbalanced data.

Feature selection. Feature selection has been widely used in healthcare to map the original feature space into a lower dimensional one. One research direction of feature extraction in healthcare is using principal component analysis (PCA) and independent component analysis (ICA) to project disease markers to another space [5]. However, the features after projection are totally new and different with original features. Although a lower dimension of features is obtained, it is hard to interpret what is the meaning of the new features by human. On the other hand, several statistical theories (e.g., Chi-Square) [6] and conventional classifiers (e.g., Logistic Regression) have been used to select significant features having higher score. These methods can be categorized as three classes [7]: filter, wrapper and embedded methods. Wrapper methods measure the significance of feature subsets based on the classifier performance. In contrast, the filter methods consider the intrinsic properties of the features (i.e., the "relevance" of the features) without incorporating any specified learning algorithm. The embedded methods select features by optimizing the objective function or performance of a learning algorithm.

However, all of the above methods select features under the assumption of linear relationship between the features and the target variable. It is important to take into account the non-linear dependency between the features and the target variable in order to select more informative features and improve the classification performance. Although deep learning combines lower-level representations to yield higher-level representations of features and model complex medical data with nonlinear structures for prediction and classification, it does not have the ability to select features directly.

Feature embedding. In the literature, many conventional deep learning models require the input to be numerical, so they convert categorical features to numerical using preprocessing, such as mapping a category to the conditional probability of a particular label [8] or one-hot encoding [9] - a method where the categorical variable is broken into as many features as the unique number of categories for that feature and for every row, 1 is assigned for the feature representing that row's category and rest of the features are marked 0. Embedding has also being used in natural language processing (NLP) for word representation. For example, a representation of each word is learned and then is fed into a neural network to make the prediction [10]. In addition to NLP, [11] proposes a data embedding method for time-related feature in the transportation field. The key idea is to embed the data into a 2-dimensional space before feature selection, and then a data-driven ensemble learning approach is applied for prediction. In the medical field, [12] presents a categorical feature embedding method to encode categorical features into vectors for autism diagnosis prediction.

However, there are a lot of issues with those methods. For categories with lots of unique features, we get a lot of sparse data using one-hot encoding. And they all require a large amount data to train the model for learning the embeddings.

Learning from few samples and imbalanced data. To learn from only few examples of each class, there are three major works in computer vision field (i.e., Face Recognition): 1) Matching Networks [13], that first embed a high dimensional sample into a low dimensional space and then perform a generalised form of nearest-neighbours classification. One of the popular matching networks is siamese networks that use a pairwise verification loss to perform metric learning and then in a separate phase use the learnt metric space to perform nearest-neighbours classification; 2) Prototypical Networks [14] that aim to learn prototypical representations; 3) Model-agnostic meta-learning [15] provides a good initialization of a model's parameters to achieve an optimal fast learning on a new task and avoids overfitting that may happen when using small data. There are two major methods to deal with imbalanced data [16, 4]: algorithm-based and sampling-based methods. For example, [17] combines cost-sensitive method with neural network architecture to deal with imbalanced medical data. [18, 19] introduce SMOTE as a resampling method to increase the number of samples in the minority class. However, we have applied the abovementioned methods on the small imbalanced medical data and experimentally concluded that none of these methods perform well to deal with imbalanced medical data with few samples.

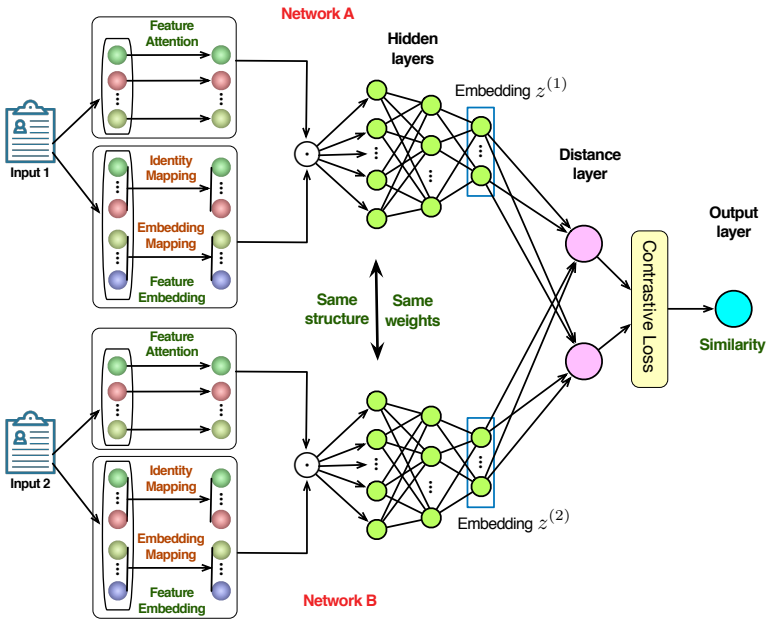


Fig. 1. The overall framework of the proposed model. An input pair is fed into the network. The input variables are first passed through the feature attention and feature embedding layer where feature attention learns feature score of each feature and feature embedding embeds categorical features and maps numerical features. The outputs from feature attention and feature embedding are concatenated by element-wise product, and then be fed into a neural network with several dense layers. The neural network learns their embeddings into constant vectors $z^{(1)}$ and $z^{(2)}$ that their similarity can be identified via a distance layer and contrastive loss, followed by an output layer where outputs the similarity of the two inputs.

III. Methods

A. Problem definition

Given a data set $\mathcal{D} = \{\mathcal{D}^{train}, \mathcal{D}^{test}\}$ for a supervised learning task, $\mathcal{D}^{train} = \{(x^{(i)}, y^{(i)})\}_{i=1}^{I}$ is a labelled training set with a small number I and skewed y distribution, and $\mathcal{D}^{test} = \{x^{test}\}$ is the test set. We aim to learn a hypothesis $\hbar$ on the training set $\mathcal{D}^{train}$, and evaluate the hypothesis on the test set. Each training example $x^{(i)} \in \mathcal{X} \subseteq \Re^d$ is represented by the learned embeddings $z^{(i)} \in \mathcal{Z} \subseteq \Re^m$, and then making prediction based on the similarity of the learned embeddings. The performance is measured by the loss function $\mathcal{L}$ between prediction $\hat{y} = \hbar(x; W)$ and true label y.

B. Model architecture

The architecture of our model is shown in Figure 1. Our model consists of three components: feature attention, feature embedding, and siamese networks. The feature attention makes the deep learning model explainable because it is able to learn feature weights and enables to select the most informative features from a large amount of input variables. The feature embedding is a compellingly simple, yet effective neural network architecture to make the model applicable to heterogeneous healthcare data, thus, the model is capable of directly handling both numerical and categorical features. The feature attention and feature embedding are in the same layer of neural networks, and their outputs are element-wise multiplied as an input to the next layer. The siamese network consists of two identical fully connected nets that take the outputs from feature attention and feature embedding after element-wise multiplication. Two inputs are fed into the two identical networks to learn their representations during training and test phases. After that, the siamese networks compare their representation vectors to output their similarity.

C. Feature attention

While deep learning has achieved great success for classification and prediction tasks, most deep learning models are not able to output which features are informative and significant to classifiers. Since the input variables may contain many redundant, noisy, or irrelevant features, not all of the variables contribute to the performance of the predictive models we build. In our proposed model, we add a feature attention layer such that helps to identify a subset of relevant input variables in a dataset.

Let $x^{(i)} \in \Re^d$ denotes the feature vector of object i, and $W_{fa} \in \Re^d$ be the feature attention (weight), such that,

$$W_{fa} = [w_{fa}^1, w_{fa}^2, \cdots, w_{fa}^d], \sum_{j=1}^{d} w_{fa}^j = 1, w_{fa}^j \geq 0. \quad (1)$$

The feature attention layer is a one-by-one connection to the input variables. It is the element-wise product between the feature attentions W_{fa} and the feature vector $x^{(i)}$ of object i. Then, we have the feature attention layer as follows:

$$h_0^{fa} = W_{fa} \odot x^{(i)} = [w_{fa}^1 x_1^{(i)}, w_{fa}^2 x_2^{(i)}, \cdots, w_{fa}^d x_d^{(i)}] \quad (2)$$

where $\odot$ is element-wise multiplication operator.

D. Feature embedding

One of the challenges to analyze healthcare data is that the data typically contains both continuous numerical variables and categorical variables. Dealing with continuous numeric

data is often easier than categorical data given that it can be fed into most of deep learning models after normalization. However, naively applying deep learning algorithms with integer representation for categorical variables does not work well. Categorical variables are known to hide and mask lots of interesting information in a dataset and they might even be the most important variables in a model. To address this problem, we introduce a feature embedding layer, capable of naturally handling both categorical and numerical features.

Given a set of inputs $\{x^{(i)}, y^{(i)}\}_{i=1}^{I}$ containing I instances. Each feature vector is a concatenation of numerical features $x_U^{(i)}$ and categorical features $x_C^{(i)}$. Our model maps the categorical variables into numerical hidden representations in Euclidean spaces, which is to build a vector embedding to every category type. The new representation will be concatenated with the numerical part and then element-wise multiplies feature attention part to be fed into the remaining neural network.

Suppose a categorical variable $x_{C_k}^{(i)}$ has p categories, and let hyperparameter η be a user defined dimension for embedding representation. We initialise an embedding matrix

$$E_C = \begin{pmatrix} [x_{C_k}^{(i)}]_1^1 & [x_{C_k}^{(i)}]_1^2 & \cdots & [x_{C_k}^{(i)}]_1^\eta \\ [x_{C_k}^{(i)}]_2^1 & [x_{C_k}^{(i)}]_2^2 & \cdots & [x_{C_k}^{(i)}]_2^\eta \\ \vdots & \ddots & \vdots & \\ [x_{C_k}^{(i)}]_p^1 & [x_{C_k}^{(i)}]_p^2 & \cdots & [x_{C_k}^{(i)}]_p^\eta \end{pmatrix} \in \Re^{p\times\eta} \quad (3)$$

where each row in the embedding matrix denotes the embedding of each category for this categorical feature, i.e., $[x_{C_k}^{(i)}]_j$ is the j^{th} category of the k^{th} categorical feature $x_{C_k}^{(i)}$.

Then we add an embedding layer in the neural network to do a lookup for a given value from the embedding matrix E_C, which returns a numerical embedding for each category. The new representation along with the numerical variables $x_U^{(i)}$ would then be fed into the next layer in the neural networks. Note that, this part is only for categorical features embedding and the embedding dimension is η while the overall embedding $z^{(i)}$ of an input $x^{(i)}$ is from the network in Section III-E with embedding dimension m.

E. Learning from few and imbalanced data

Deep neural networks have been successfully used for classification tasks on large datasets by minimizing a cross-entropy loss function. However, there are many problems (i.e., overfitting) if we train a neural network classifier based on cross-entropy on a small or an imbalanced dataset, because there are tens of thousands parameters need to be optimized. Therefore, in this paper, we use metric learning which is a non-parametric method to calculate the distance between two samples based on their learned embeddings.

We use a siamese network to learn from few and imbalanced data. The siamese network consists of two channels where use two identical dense networks. The input to the siamese networks should be in pairs, along with their labels, stating whether the input pairs are a genuine pair (same) or an opposite pair (different). During training, the input pairs are constructed by the way of combining two samples from the training data. Then input pairs are first fed into the first layer that consists of feature attention and feature embedding. Outputs from feature attention and feature embedding are concatenated by element-wise multiplication operation, and then be fed into several dense layers with $ReLU$ non-linear activation for embedding learning. The embedding learning embeds $x^{(i)} \in \mathcal{X} \subseteq \Re^d$ to a smaller embedding space $z^{(i)} \in \mathcal{Z} \subseteq \Re^m$, where the similar and dissimilar pairs are able to compute and identify. We then feed these embeddings to an energy function (i.e., Euclidean distance, Cosine similarity, and Manhattan distance) which will give us a similarity between the two inputs. The value from the energy function will be smaller if the two inputs are similar (i.e., they are from the same class). Otherwise, the value will be larger.

For example, let's say we have two inputs $x^{(1)}$ and $x^{(2)}$. As shown in Figure 1, we feed $x^{(1)}$ to network A and $x^{(2)}$ to network B. The role of both of these networks is to generate embeddings (feature vectors) of the inputs. We can use any network architecture (i.e., CNN or NN) that is able to give us embeddings. Then we will feed these embeddings to the distance layer which tells us how similar the two inputs are. For the distance layer, we can use any similarity measure such as Euclidean distance, Cosine similarity, Manhattan distance, and so on. Since networks A and B are identical (share weights and same architecture), the learned embeddings $z^{(1)}$ and $z^{(2)}$ should be similar if the two inputs are from the same class and dissimilar if they are from different classes.

F. Loss function

Now the problem turns to how to train the proposed network. Let $z^{(i)}, z^{(j)} \in \mathcal{Z}$ be a pair of learned embeddings of inputs $x^{(i)}$ and $x^{(j)}$. We use $\Psi = 1$ when the inputs $x^{(i)}$ and $x^{(j)}$ are from the same class, and $\Psi = 0$ otherwise. A siamese network is trained with the learned embeddings being fed to a contrastive loss. We impose feature attention by introducing a regularization term in a loss function. Thus, our objective function is as follows:

$$\begin{aligned}\mathcal{L}(x^{(i)},x^{(j)},\Psi) = \frac{1}{2}(1-t)(D(x^{(i)},x^{(j)}))^2 + \\ \frac{1}{2}\Psi(max\{0,\epsilon - D(x^{(i)},x^{(j)})\})^2 + \lambda\|\frac{W_{fa}}{d}\|\end{aligned} \quad (4)$$

where $\lambda \in [0,1]$ is a user-specified parameter coefficient, and d is a number of variables in a dataset. W_{fa} is the weights of feature attention layer. $D(x^{(i)}, x^{(j)})$ is the parameterized cosine similarity between the learned embeddings of $z^{(i)}$ and $z^{(j)}$. That is

$$D(x^{(i)}, x^{(j)}) = \frac{z^{(i)}z^{(j)}}{||z^{(i)}|| \cdot ||z^{(j)}||} \quad (5)$$

Note that $D(x^{(i)}, x^{(j)})$ could be any similarity measure such as Euclidean distance, Cosine similarity, and so on. But the reason why we use Cosine similarity instead of Euclidean distance in this paper is to make our model also work well to the difference in length of the learned embeddings. So the loss will decrease $D(x^{(i)}, x^{(j)})$ when the samples are from the same class, on the other hand, when they are dissimilar it will try to increase $D(x^{(i)}, x^{(j)})$ with a certain margin $\epsilon(\epsilon > 0)$.

TABLE I
DATASETS USED IN THE STUDY

Data Sets	# of training samples (positive / negative)	# of test samples (positive / negative)
MIMIC-III AKI Data	887 / 53	591 / 35
eICU AKI Data	1335 / 312	667 / 156
Framingham Heart Study	1078 / 193	359 / 64

ϵ is a margin value which is greater than 0. The term margin is used to hold the constraint, that is, when two input values are dissimilar and if their distance is beyond this margin, then they do not incur a loss. $max()$ is a function denoting the bigger value between 0 and $\epsilon - D(x^{(i)}, x^{(j)})$.

$\mathcal{L}(x^{(i)}, x^{(j)}, \Psi)$ can be minimized with gradient descent optimization, and the standard back-propagation algorithm can be applied.

IV. RESULTS

We compare our model with four baselines that used for few or imbalanced data analysis on one synthetic data and three real-world medical data.

A. Simulations

To show the ability of the proposed model is able to detect key and noisy features, we generate a synthetic data with Gaussian noise as simulation following a method in [20]. We sampled 1500 data points from each of the two 2-dimensional Gaussian distributions with the same covariance matrix $\sum = diag(1, 1)$. The true signal is set by two different mean values $\mu = (0, 0)$ and $(0, 5)$. Then, we added 40-dimensional Gaussian noise with mean $\mu = (2.5, 2.5, \cdots 2.5)$ and covariance $\sum = diag(10, 10, \cdots, 10)$, resulting each point has 42 dimensions and the first two are the true label while the rest being random noise. Figures 2 and 3 illustrate the data points with the true signal (the first two dimensions) and the "corrupted" signal (using tSNE for the 42-dimensional vector) respectively.

We randomly select 70% data to train our model IDEM and a Random Forest classifier, and compare the learned feature weights from the two models. Figure 4 demonstrates the learned weights by IDEM and Random Forest (RF) for the 42-dimensional input features, where we can observe that IDEM model learns the weights of the true signal are much higher than those in the noise. However, RF is not able to distinguish the true signal and noise very well. Both models are able to output the score of each feature, which is to distinguish the feature importance. However, the ability of distinguishing feature importance is subject to the performance of classification. Since IDEM outperforms RF in terms of all criteria (e.g., Specificity and Sensitivity) on the imbalanced datasets, IDEM has superior ability of identifying real importance of features.

B. Datasets

We carry out a set of experiments to verify the effectiveness of the proposed framework on three publicly available real imbalanced medical datasets with only few samples. Table I shows the number of positive and negative samples in both training and test sets.

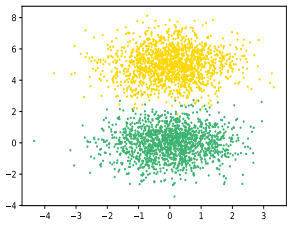


Fig. 2. Clustering data points generated by simulations.

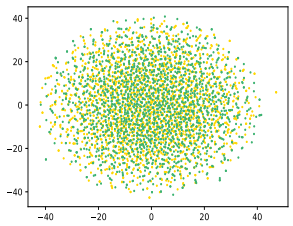
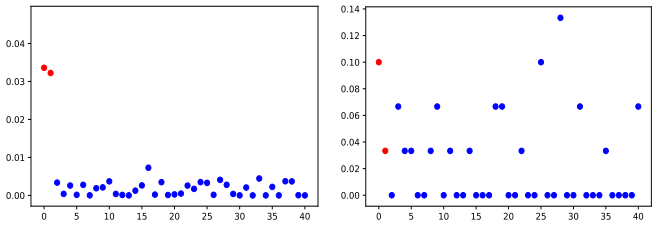


Fig. 3. After adding 40-dimensional Gaussian noise.



(a) Feature weights from IDEM. (b) Feature weights from RF.

Fig. 4. Learned feature weights by IDEM and RF. The red dots denote the true signals and the blue dots corresponding to noise.

- **MIMIC-III AKI Data**. This is a publically available Electronic Health Record (EHR) data sets and Kidney Disease Improving Global Outcomes (KDIGO) criterion to definite Acute Kidney Injury (AKI). The aim is to precisely predict whether a certain patient will suffer from AKI after admission in ICU according to the last measurements of the 16 blood gas and demographic features.
- **eICU AKI Data**. This is another publically available EHR data for kidney disease. This dataset is also used for AKI prediction, and it is from the eICU database. There are more samples available in this dataset than in the MIMIC-III AKI data.
- **Framingham Heart Study Data**. This dataset is from an ongoing cardiovascular study on residents of the town of Framingham in Massachusetts, and it is publically available on the Kaggle website. The classification goal is to predict whether the patient has 10-year risk of future coronary heart disease (CHD). The dataset provides the patients' information that includes 15 attributes.

C. Benchmark methods

Since our method is efficient to learn from few and imbalanced data, we compare our model with baselines that are used for few-shot learning and imbalanced data analysis.

- **Siamese Neural Network with Feature Attention (SNN-FA)**. Siamese Neural Network [21] is one of the most prominent deep neural networks for few-shot learning. To evaluate the effectiveness of using embedding for categorical feature in the first layer of the proposed architecture, we compare the siamese neural network by adding only feature attention (without the feature embedding) in the first layer as a baseline.
- **K-Nearest Neighbours (KNN)** [22] is the simplest way of doing classification. KNN calculates the Euclidean distance of the test sample $\hat{x}$ from each training example and picks the closest one as follows:

$$C(\hat{x}) = \underset{c \in D^{train}}{\arg\min} ||\hat{x} - x_c|| \tag{6}$$

  where x_c is the training example in category c.
- **Cost-Sensitive Deep Neural Network (CSDNN)** [17] considers costs of misclassified instances that vary by type of category whereas traditional classification models assume that all misclassification errors carry the same cost. CSDNN does not directly create a balanced class distribution. Instead, it highlights the imbalanced learning problem using a cost matrix that describes the cost of misclassification in a particular scenario. The cost-sensitive loss function is

$$\begin{aligned}\mathcal{L}_{CS} = y * (c_{FN} * log(\hat{y}) + c_{TP} * log(1-\hat{y})) + \\ (1-y) * (c_{FP} * log(1-\hat{y}) + c_{TN} * log(\hat{y}))\end{aligned} \tag{7}$$

  where $c_{FN}, c_{TP}, c_{FP}, c_{TN}$ represent the costs of false negative, true positive, false positive, and true negative. We use [2,1,18,1], [2,1,6,1], and [2,1,6,1] for MIMIC-III, eICU, and Framingham Heart data, respectively.
- **The Synthetic Minority Oversampling Technique (SMOTE)** [19] is the most widely used algorithm to balance healthcare data. SMOTE is a more advanced oversampling method which interpolates among existing minority class examples and generates new minority class samples. After the over-sampling process, a Random Forest classifier is applied to classify the new balanced dataset.
- **RUSBoost** [23] is a variant of AdaBoost to handle the imbalance problem. AdaBoost is one of Boosting algorithms, which trains several weak classifiers using a same dataset. RUSBoost uses random under-sampling integrated in the learning of AdaBoost. During learning, the problem of class balancing is alleviated by random under-sampling the sample at each iteration of the boosting algorithm.
- **Cost-Sensitive SVM (CS_SVM)** [24] is a modification of SVM that weighs the margin proportional to the class importance.

D. Performance evaluation criteria

Appropriate evaluation criteria are crucial for assessing the performance. In the case of imbalanced classes, accuracy metric can be misguiding, because high metric does not show prediction capacity for the minority class. As the minority class may bias the decision boundary and has little impact on accuracy, we use evaluation criteria include Specificity, Sensitivity, F1 Score, ROC (Receiver Operating Characteristic) curve, and the positive predicted value (PPV) and negative predicted value (NPV) based on confusion matrix.

In the context of imbalanced classification, sensitivity is the percentage of correctly classified minority instances. In contrast, specificity denotes the percentage of correctly classified majority instances. PPV denotes the percentage of relevant objects that are identified for retrieval. F1 score represents a harmonic mean between specificity and PPV.

E. Results and discussion

Results: Table II presents the performance of the proposed model IDEM and the baselines on the various datasets. In comparison to the state-of-the-art baselines on the test data, we observe that our model (IDEM) performs better than baselines in terms of Specificity, Sensitivity, PPV, NPV and F1 Score. For example, all the methods have relatively high specificity on the Framingham Heart Study Data. But for the SMOTE, KNN and CSDNN classifiers, the sensitivity values are very pool (25.2%, 24.1% and 31.4%). Obviously, the sensitivity and PPV are significantly improved by using the proposed framework. Also, the proposed method achieves superb F1 Score on the three datasets compared to other baselines. The sensitivity and F1 Score of IDEM are 2 to 7 times higher than using SMOTE, KNN and CSDNN. As we can observe from the results, under the same false positive rate, we are able to predict kidney or heart disease with high true positive rate, which is much better than that of in baselines.

To further validate the interpretability of IDEM, we add two new variables into the MIMIC-III AKI data. The first feature $Y\_correlated$ is highly correlated to the target and the second X_random is a random noise feature generated from a normal distribution. From Figure 5, our model detects the highly correlated feature with the highest score while gives the noisy feature a very low score.

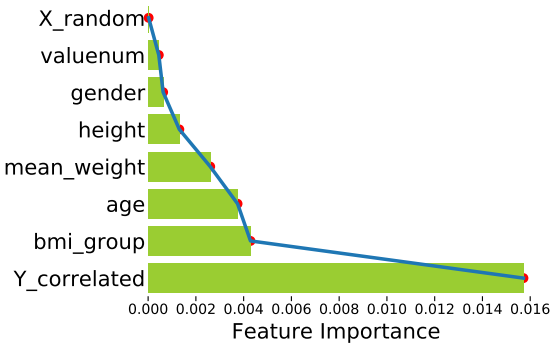


Fig. 5. Feature importance analysis for the MIMIC-III AKI data.

TABLE II
EXPERIMENTAL RESULTS

Datasets	Methods	Specificity	Sensitivity	PPV	NPV	F1 Score	AUC
MIMIC-III AKI	SMOTE	0.973	0.122	0.657	0.721	0.206	0.714
	RUSBoost	0.958	0.155	0.349	0.887	0.860	0.668
	CS_SVM	0.882	0.387	0.164	0.960	0.850	0.655
	KNN	0.948	0.375	0.086	0.991	0.139	0.670
	CSDNN	0.963	0.091	0.571	0.660	0.156	0.616
	SNN-FA	0.987	0.900	0.771	0.994	0.831	0.883
	IDEM	**0.992**	**0.938**	**0.857**	**0.997**	**0.896**	**0.927**
eICU AKI	SMOTE	0.847	0.239	0.539	0.598	0.331	0.575
	RUSBoost	0.834	0.235	0.429	0.673	0.630	0.571
	CS_SVM	0.846	0.269	0.290	0.832	0.649	0.567
	KNN	0.812	0.205	0.103	**0.907**	0.137	0.572
	CSDNN	0.857	0.228	**0.660**	0.477	0.339	0.564
	SNN-FA	0.863	0.515	0.587	0.846	0.633	0.657
	IDEM	**0.869**	**0.536**	0.611	0.857	**0.652**	**0.694**
Framingham Heart Study	SMOTE	0.896	0.252	0.531	0.719	0.342	0.689
	RUSBoost	0.877	0.221	0.226	0.919	0.330	0.583
	CS_SVM	0.801	0.396	**0.807**	0.192	0.460	0.599
	KNN	0.855	0.241	0.109	0.904	0.151	0.582
	CSDNN	0.890	0.314	0.649	0.557	0.321	0.619
	SNN-FA	0.893	0.711	0.649	0.901	0.676	0.688
	IDEM	**0.898**	**0.775**	0.796	**0.927**	**0.791**	**0.719**

Discussion: The specificity is high using all the methods because the classifier was able to classify the majority class (positive) samples well but failed in classifying the minority ones. This is critical because the misclassification cost of cases is more serious in healthcare. Therefore, rather than only considering the classification accuracy, other criteria (i.e., Sensitivity, PPV and NPV) are also important to measure the performance for imbalanced datasets. As shown in Table II, the proposed IDEM outperforms other baselines in terms of sensitivity, PPV and NPV, which indicates IDEM can better handle imbalanced data. SMOTE and KNN achieve a low sensitivity and PPV. This is because SMOTE and KNN do not pay enough attention to the minority class. SNN-FA outperforms other baselines in terms of sensitivity and PPV, which indicates the siamese network framework is a good solution to analyze the imbalanced data. It is observed from the experiments that the majority and minority ratio is not the only issue in building a good predictive model. There is also a need for enough training samples that display data properties consistent with the class label assigned to them. The SNN-FA method performs better than conventional machine learning and deep neural networks because we only have few samples for training and the siamese network learns better from few samples. Since our model leverages both the siamese network and deep feature embedding, IDEM outperforms other baselines on the few and imbalanced datasets. The feature attention layer in the IDEM makes it interpretable to identify which features contribute to the classifier in distinguishing case and control.

V. CONCLUSION

Few samples and class imbalance are common problems with most medical datasets [25]. Although deep learning has achieved great success in classifying clinical diagnosis [26], it requires plenty of training data. Most existing deep learning methods tend to overfit and fail to generalize in biomedical research where the training sample size is usually small and unbalanced. To address this problem, we propose the IDEM model that contains feature attention and feature embedding using siamese networks. We conducted extensive experiments on three medical datasets with few samples and highly skewed class distributions. From the results, IDEM is able to identify noise features and significant features. Compare with traditional methods, IDEM is more effective to classify the minority samples, and it exhibits obvious advantages when the dataset is extremely imbalanced.

REFERENCES

[1] A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, C. Cui, G. Corrado, S. Thrun, and J. Dean, "A guide to deep learning in healthcare," *Nature medicine*, vol. 25, no. 1, pp. 24–29, 2019.

[2] T. Fu, T. N. Hoang, C. Xiao, and J. Sun, "Ddl: deep dictionary learning for predictive phenotyping," in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, 2019.

[3] H. Afrabandpey, T. Peltola, and S. Kaski, "Human-in-the-loop active covariance learning for improving prediction in small data sets," *Proceedings of International Joint Conference on Artificial Intelligence*, 2019.

[4] Y. Zhao, Z. S.-Y. Wong, and K. L. Tsui, "A framework of rebalancing imbalanced healthcare data for rare events' classification: a case of look-alike sound-alike mix-up incident detection," *Journal of healthcare engineering*, vol. 2018, 2018.

[5] L. B. Frazao, N. Theera-Umpon, and S. Auephanwiriyakul, "Diagnosis of diabetic retinopathy based on holistic texture and local retinal features," *Information Sciences*, vol. 475, pp. 44–66, 2019.

[6] H. Wang and P. Avillach, "Diagnostic classification and prognostic prediction using common genetic variants in autism spectrum disorder: Genotype-based deep learning," *JMIR medical informatics*, vol. 9, no. 4, p. e24754, 2021.

[7] J. Li and H. Liu, "Challenges of feature selection for big data analytics," *IEEE Intelligent Systems*, vol. 32, no. 2, pp. 9–15, 2017.

[8] W. Chen, Y. Chen, and K. Q. Weinberger, "Fast flux discriminant for large-scale sparse nonlinear classification," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2014, pp. 621–630.

[9] C.-W. Hsu, C.-C. Chang, C.-J. Lin *et al.*, "A practical guide to support vector classification," 2003.

[10] D. Ganguly, D. Roy, M. Mitra, and G. J. Jones, "Word embedding based generalized language model for information retrieval," in *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*. ACM, 2015, pp. 795–798.

[11] X. Zhang, Z. Zhao, Y. Zheng, and J. Li, "Prediction of taxi destinations using a novel data embedding method and ensemble learning," *IEEE Transactions on Intelligent Transportation Systems*, 2019.

[12] H. Wang, L. Li, L. Chi, and Z. Zhao, "Autism screening using deep embedding representation," in *International Conference on Computational Science*. Springer, 2019, pp. 160–173.

[13] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra *et al.*, "Matching networks for one shot learning," in *Advances in neural information processing systems*, 2016, pp. 3630–3638.

[14] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Advances in Neural Information Processing Systems*, 2017, pp. 4077–4087.

[15] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 1126–1135.

[16] S. H. Dumpala, R. Chakraborty, S. K. Kopparapu, and T. Reseach, "A novel data representation for effective learning in class imbalanced scenarios." in *IJCAI*, 2018, pp. 2100–2106.

[17] H. Wang, Z. Cui, Y. Chen, M. Avidan, A. B. Abdallah, and A. Kronzer, "Predicting hospital readmission via cost-sensitive deep learning," *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, vol. 15, no. 6, pp. 1968–1978, 2018.

[18] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[19] A. Fernández, S. Garcia *et al.*, "Smote for learning from imbalanced data: progress and challenges, marking the 15-year anniversary," *Journal of artificial intelligence research*, vol. 61, pp. 863–905, 2018.

[20] T. Ma and A. Zhang, "Affinitynet: semi-supervised few-shot learning for disease type prediction," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 1069–1076.

[21] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," in *ICML deep learning workshop*, vol. 2, 2015.

[22] P. Cunningham and S. J. Delany, "k-nearest neighbour classifiers," *Multiple Classifier Systems*, vol. 34, no. 8, pp. 1–17, 2007.

[23] C. Seiffert, T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano, "Rusboost: A hybrid approach to alleviating class imbalance," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 40, no. 1, pp. 185–197, 2009.

[24] A. Iranmehr, H. Masnadi-Shirazi, and N. Vasconcelos, "Cost-sensitive support vector machines," *Neurocomputing*, vol. 343, pp. 50–64, 2019.

[25] J. M. Johnson, "Survey on deep learning with class imbalance," *Journal of Big Data*, vol. 6, no. 1, 2019.

[26] B. Norgeot, B. S. Glicksberg, and A. J. Butte, "A call for deep-learning healthcare," *Nature medicine*, vol. 25, no. 1, p. 14, 2019.