# Deep Learning within Tabular Data: Foundations, Challenges, Advances and Future Directions

Weijieying Ren<sup>1</sup>, Tianxiang Zhao<sup>1</sup>, Yuqing Huang<sup>2</sup> and Vasant Honavar<sup>1</sup>

<sup>1</sup>Information Sciences and Technology, The Pennsylvania State University <sup>2</sup>Computer Science and Technology, University of Science and Technology of China {wjr5337, tkz5084, vuh14}@psu.edu, enthlinn@mail.ustc.edu.cn

#### 1 Introduction

# 1.1 Related Surveys

Based on the background discussed above, designing a state-of-the-art representation learning method for tabular data involves three fundamental elements: training data, network architectures, and learning objectives. To improve both the quantity and quality of training data, various data-related techniques, such as data augmentation and generation, are employed or introduced. To better leverage the inherent properties of tabular data, the neural architectures are specifically designed to capture: 1) irregular patterns within each column, including variations in scale, mean, and variance. and 2) complex inter-relationships among different columns. Finally, multiple learning strategies and objectives are defined to facilitate the learning of high-quality representations.

Despite incorporating three key design elements, most existing surveys on tabular data representation learning primarily concentrate on either neural architectural features or learning methodologies. An early survey article [Sahakyan et al., 2021] provides a comprehensive overview of Explainable Artificial Intelligence (XAI) techniques applicable to tabular data, with a particular emphasis on feature transformation and classical machine learning methods in classification tasks. Subsequently, several surveys have explored synthetic data generation methods. instance, [Sauber-Cole and Khoshgoftaar, 2022] reviews research on data generation within the healthcare domain, specifically focusing on GAN-based techniques. In addition, [Sauber-Cole and Khoshgoftaar, 2022] investigates the application of GANs to address the class imbalance problem. Among these works, [Borisov et al., 2022a] presents a comprehensive survey that discusses feature transformation, neural network design, and data generation challenges within a broader context. In contrast, [Wang et al., 2024a] systematically reviews and summarizes the recent advancements and challenges associated with self-supervised learning for non-sequential tabular data [Ren and Honavar, 2024]. Furthermore, with the emergence of foundation models and transformer-based large language models (LLMs), recent articles [Ruan et al., 2024] examine the adaptation of these models to tabular data, concentrating primarily on their learning aspects.

Unlike previous works, we present a comprehensive review

of representation learning methods for tabular data, focusing on their universality—effectiveness across a range of downstream tasks. Our discussion provides insights into the underlying intuitions driving these methods and examines how they enhance the quality of learned representations across all three key design aspects. Specifically, we aim to identify and analyze research directions informed by recent state-of-theart studies that concentrate on neural architecture design, the formulation of corresponding learning objectives, and the effective utilization of training data to improve the quality of learned representations for various downstream tasks. Table 1 highlights the distinctions between our survey and existing literature.

### 1.2 Survey Scope and Literature Collection

For literature review, we use the following keywords and inclusion criteria to collect literatures.

**Keywords**. "tabular" or "'table", "tabular" AND "'representation", "tabular" AND "'embedding", "tabular" AND "'modeling", "transaction data" AND "representation", "biomedical data" AND "representation". We use these keywords to search well-known repositories, including ACM Digital Library, IEEE Xplore, Google Scholars, Semantic Scholars, and DBLP, for the relevant papers.

**Inclusion Criteria** Related literatures found by the above keywords are further filtered by the following criterion. Only papers meeting these criteria are included for review.

- Written exclusively in the English language
- Focused on approaches based on deep learning or neural networks
- Published in or after 2020 in reputable conferences or high-impact journals

Quantitative Summary Given the above keywords and inclusion criteria, we selected 127 papers in total. Fig. 3 shows the quantitative summary of the paper selected for review. We can notice from Fig. 3a that neural architectures and learning objectives are similarly considered important in designing state-of-the-art methods. Most papers were published at ICLR, NeurIPS, followed by AAAI and ICML (Fig. 3b). According to Fig. 3c, we expect more papers on this topic will be published in the future.

# 2 Preliminary

This section provides definitions and notations used in the paper, describes downstream tasks for tabular data analysis, and highlights the unique properties of tabular data.

#### 2.1 Definitions

**Definition 1.** (Tabular Data). Tabular data is systematically arranged in a structured format characterized by rows and columns. Each row denotes an individual sample record, while each column signifies a distinct type of feature observation. Each column is composed of a header and a series of values, commonly referred to as cells. Each row is represented as a vector of  $M_{num}$  numerical features and  $M_{cat}$  categorical features  $\mathbf{x} = [\{\mathbf{x}_{num}^i\}_{i=1}^{M_{num}}, \{\mathbf{x}_{cat}^i\}_{i=1}^{M_{cat}}]$ , where  $\mathbf{x}_{num}^i \in R$  and  $\mathbf{x}_{cat}^i \in \{1, 2, ..., C_i\}$ .  $C_i$  denotes the size of finite candidate values for the i-th categorical feature.

**Definition 2.** (Classification). Tabular data classification aims to assign predefined class labels  $Y = \{y_1, y_2, ..., y_C\}$  to each row of tabular data. Denoted D as a tabular dataset with N samples, X is a row of tabular data and y is the corresponding labels. where  $y_i \in \{-1, 1\}$  is a binary classification task and  $y_i \in \{1, 2, ..., C\}$  is a multi-class classification task.

**Definition 3.** (Regression). Tabular data regression has a similar goal to classification tasks, with a key difference in label annotation. In tabular data regression, the objective is to predict a continuous value  $y \in R$  for each row.

**Definition 4.** (Clustering). Tabular data clustering aims to partition X into a group of clusters  $G = g_1, g_2, ..., g_G$  by maximizing the similarities between tabular rows within the same cluster and the dissimilarities between tabular rows of different clusters.

**Definition 5.** (Anomaly Detection). Tabular anomaly detection is a crucial process for identifying tabular samples within a dataset that significantly deviate from established patterns of normal behavior. This approach typically involves training a model on a labeled dataset D, where the model learns the characteristics that define normal observations. Once trained, the model computes anomaly scores  $A = (a_i, ..., a_{|X|})$  for each row in an unseen test set  $X_{test}$ . These scores quantify the degree of deviation for each observation. The final classification of anomalies is made by comparing each score  $a_i$  against a predetermined threshold  $\delta$ : an sample is classified as anomalous if  $a_i > \delta$  and as normal otherwise.

**Definition 6.** (Imputation of Missing Values). Tabular imputation aims to fill missing values with plausible values to facilitate subsequent analysis. Given tabular data X and known binary matrix  $M \in R$ ,  $x_i$  is missing if  $m_i = 0$ , and is observed otherwise. The imputed tabular data is given as:  $X_{imputed} = X \odot M + \hat{X} \odot (1 - M)$ .

**Definition 7.** (Retrieval). Tabular retrieval aims to obtain a set of samples that are most similar to a query provided. Given a query sample X and a similarity measure  $f(\cdot)$ , find an ordered list  $Q = \{X_{i=1}^k\}$  of tabular samples in the given dataset or database, containing tabular samples that are the most similar to tabular queries.

**Definition 8.** (Test Time Adaptation on Tabular Data). Given a pre-trained source model  $f(\cdot)$ , TTA adapts source domain model parameters  $\theta$  to obtain target domain parameters  $\theta'$  using unlabeled target domain data DT = (xi)NTi = 1. It should be noted that the feature spaces for the source and target domains are identical; however their distributions are not

# 2.2 Foundational Properties In Tabular data Modeling

In this subsection, we elaborate on the distinctive characteristics of tabular data and corresponding distinct research perspectives. Due to these specific properties, methodologies developed for image or text data often unsuitable to be applied directly to tabular data.

Heterogeneity Tabular data exhibits heterogeneity due to the incorporation of various data types, including discrete entities such as categorical and binary features, continuous variables represented by numerical data. Furthermore, columns with the same data type may still display distinct marginal distributions, as evidenced by differences in statistical properties such as mean, variance, and scale.

Diverse and Contextualized Semantic Meaning Tabular data exhibits diverse semantic meanings among its columns. For instance, in a clinical diagnosis prediction task, the value 100 is ambiguous and lacks meaning until contextualized, such as by specifying 100 kg or 100 ml. In addition, the marginal distribution of features can vary across different contextual settings, even if they share the same semantic meaning. For example, in a weather prediction task, the statistical characteristics of air moisture may differ significantly between eastern and western regions. This property complicates the transfer of knowledge across domains and tasks, and it also poses challenges for tabular imputation, often requiring the expertise of domain specialists.

Permutation Invariance and Equivalence Permutation Invariance refers to the property that the outcomes of an analysis or model remain unchanged when the rows or columns of a tabular dataset are permuted. This means that rearranging the order of the observations (rows) or the features (columns) does not affect the statistical properties, relationships, or predictions derived from the data. Permutation Equivalence indicates the tabular dataset remains essentially the same in terms of its statistical properties or relationships, regardless of the order of observations or features. Consequently, the results of analyses, such as predictions or statistical measures, remain consistent even when specific operations, such as normalization or scaling of features, are applied.

High Noise and Missing Value Tabular data, especially in real-world environments, often contains noise and missing values. This noise typically arises from measurement errors or annotation mistakes. Missing values can also stem from measurement errors and may exhibit random missing patterns. In some cases, missing values can convey important information and indicate "missing not at random" patterns. For example, in electronic health record data, each patient may undergo only specific lab tests that are critical for accurate diagnosis verification. These partial observations actually provide essential information for a model to understand

the relationships between biomedical features and their corresponding diagnosis labels.

Inter-Relationship across Columns or Rows In tabular data, both columns and rows reveal associations. To model column-wise interactions, existing works analyze this property mainly from two directions: (1) Statistical Dependency Different columns may have relevance with each other due to overlapped semantic information. For example, 'age' and 'date of birth' share statistical overlap when recording patient demographic information. Consequently, it is desired to exploit such dependency during encoding and modeling tabular data distribution, (2) Latent Relations There would be latent causal factors behind multiple column observations. With this observation, related columns or abstract concepts can be integrated and processed collectively to extract such critical latent factors. For instance, the family history of disease exhibits a distinct semantic correspondence to diabetes. Hence, the critical factor behind the user's vulnerability can be reflected by collecting and aggregating columns regarding family disease histories. Meanwhile, summarizing grouplevel behaviors and modeling the corresponding relationships among samples are essential for enhancing model prediction accuracy. For instance, various medical settings are characterized by the presence of multiple distinct patient behaviors. Identifying and characterizing these subgroups is crucial for understanding underlying diseases and improving the delivery of medical care.

Particularly, researches have been made against heterogeneity the following dimensions: (1) Modeling with different distributions for encoding columns of different data formats or frequencies of elements. (2) Separately learning on different columns with metrics that are sound w.r.t information theory. The basic assumption is that disparities exist across information contained in each columns. (3) Designing special element-level representation units for each column. For example in a continuous column, its numerical distribution range can be partitioned into distinct bins, with each data value being transformed into the index of corresponding interval bin. This transformation can map the raw input into a more interpretable and robust representation space.

#### 2.3 Applications

**HealthCare** Healthcare records employ tabular data structures for the storage and management of comprehensive patient health information [Ma *et al.*, 2022a; Ma *et al.*, 2022b; Ma *et al.*, 2023]. Each column corresponds to specific data types, encompassing personal details, medical images [Liu *et al.*, 2017; Li *et al.*, 2019], medical history, and diagnostic and treatment information. The utilization of tabular data in healthcare tasks leads to various classifications:

- Patient Outcome Prediction. Examples include patient mortality prediction, predicting the diagnosis of diseases, and forecasting patient responses to specific drugs.
- Clinical Trial Outcome Prediction. This involves predicting the likelihood of a clinical trial succeeding in obtaining approval for commercialization.
- Tabular Search and Retrieval Problems. Tasks under

- this category encompass clinical trial retrieval, where relevant trials are identified based on a given query or input trial; and insurance retrieval, involving the search for patient historical information.
- Tabular Generation. An example is Trial Patient Simulation, where the generation of synthetic clinical trial patient records facilitates data sharing across institutes while safeguarding patient privacy.
- Tabular Data Transference Tabular data transfer is fundamental to Health Information Exchange initiatives, where different healthcare entities share patient data securely. This promotes coordinated care, supports multidisciplinary care, reduces duplicate tests, and enhances overall healthcare efficiency.

**E-commerce** In e-commerce, tabular data is widely used for various purposes to organize, analyze, and present information related to products, transactions, and customer interactions. The industrial practice of tabular data can be summarized into the following categories:

- Product Recommendations. E-commerce platforms use tabular data to generate personalized product recommendations. Columns include customer inquiries, support tickets, resolutions, customer browsing and purchase history.
- Transaction Fraud Detection. Tabular data is utilized for monitoring transactions and detecting fraudulent activities. Columns may include transaction details, payment methods, and fraud indicators.
- **Product Search**. E-commerce platforms employ tabular search and retrieval for product searches. Users can search for products based on various criteria, such as category, price range, and specifications, and the system retrieves matching products.
- Tabular Transference. Transfer learning allows ecommerce businesses to leverage existing knowledge from one context to improve performance in related tasks, reducing the need for extensive training on new datasets. It can lead to more efficient model training, faster adaptation to new markets Real-world application relates to Supply Chain Optimization, Dynamic Pricing prediction and Customer Lifetime Value Prediction.

Energy Management Energy Management: Utilities and energy companies use tabular data to monitor energy consumption, analyze usage patterns, and optimize distribution networks for efficient energy management. Realworld application relates to peak demand prediction, Renewable Energy Source Prediction, Carbon Emission Data Retrieval.

# 2.4 Foundational Properties In Tabular data Modeling

Learning from tabular data poses a significant challenge due to the heterogeneous column contents. Different columns usually contain distinct semantics and display diverse distributions, leading to difficulties in learning the representation space. Furthermore, latent relations often exist behind observed columns, and modeling such interactions could be important for discovering critical factors. Extensive researches

have been conducted in this domain, and in this section we elaborate on these properties and distinct research perspectives on top of them.

Heterogeneous Tabular data exhibits heterogeneity due to the inclusion of diverse data types, encompassing discrete entities such as categorical and binary features, as well as continuous variables like numerical features. Meanwhile, columns sharing identical data types may still manifest distinct marginal distributions, exemplified by variations in statistical properties such as mean and variance even when both following the Gaussian distribution. Particularly, researches have been made against heterogeneity the following dimensions: (1) Modeling with different distributions for encoding columns of different data formats or frequencies of elements. (2) Separately learning on different columns with metrics that are sound w.r.t information theory. The basic assumption is that disparities exist across information contained in each columns. (3) Designing special element-level representation units for each column. For example in a continuous column, its numerical distribution range can be partitioned into distinct bins, with each data value being transformed into the index of corresponding interval bin. This transformation can map the raw input into a more interpretable and robust representation space.

**Interaction** In tabular data, columns (or abstract concepts) frequently have relations with other. Essentially, existing works analyze this property mainly from two directions:

- Statistical Dependency Different columns may have relevance with each other due to overlapped semantic information. Consequently, it is desired to exploit such dependency during encoding and modeling tabular data distribution.
- Collective Encoding There would be latent causal factors behind multiple column observations. With this observation, related columns or abstract concepts can be integrated and processed collectively to extract such critical latent factors. For instance, the family history of disease exhibits a distinct semantic correspondence to diabetes. Hence, the critical factor behind the user's vulnerability can be reflected by collecting and aggregating columns regarding family disease histories.

Explorations against column-wise interaction can be categorized based on the type of assumed interaction information and the interaction mechanism. Consideration points include the extent and dimensions in which information can be shared across columns (feature-level) and rows (sample-level), and the form of statistical association such as linear or non-linear formulations. Interaction modeling is also important for the missing value problem. In many realistic tasks, tabular data often suffers from incompleteness due to data privacy or mistakes in data collection, including biostatistics [Mackinnon, 2010], finance, algriture, etc. Column-wise interaction modeling is critical for the quality of imputation against missing values.

# 2.5 Applications

HealthCare Healthcare records employ tabular data structures for the storage and management of comprehensive pa-

tient health information. Each column corresponds to specific data types, encompassing personal details, medical history, and diagnostic and treatment information. The utilization of tabular data in healthcare tasks leads to various classifications: (I) Patient Outcome Prediction. Examples include patient mortality prediction, predicting the diagnosis of diseases [Liang et al., 2024], forecasting patient responses to specific drugs and to depression [Qin et al., 2023; Zhao et al., 2023]. (II) Clinical Trial Outcome Prediction. This involves predicting the likelihood of a clinical trial succeeding in obtaining approval for commercialization. (III) Tabular Search and Retrieval Problems. Tasks under this category encompass clinical trial retrieval, where relevant trials are identified based on a given query or input trial; and insurance retrieval, involving the search for patient historical information. (IV) Tabular Generation. An example is Trial Patient Simulation, where the generation of synthetic clinical trial patient records facilitates data sharing across institutes while safeguarding patient privacy. (V) Tabular Data Transference Tabular data transfer is fundamental to Health Information Exchange initiatives, where different healthcare entities share patient data securely. This promotes coordinated care, supports multidisciplinary care, reduces duplicate tests, and enhances overall healthcare efficiency.

**E-commerce** In e-commerce, tabular data is widely used for various purposes to organize, analyze, and present information related to products, transactions, and customer interactions. The industrial practice of tabular data can be summarized into the following categories: (I) Product Recommendations. E-commerce platforms use tabular data to generate personalized product recommendations. Columns include customer inquiries, support tickets, resolutions, customer browsing and purchase history. (II) Transaction Fraud Detection. Tabular data is utilized for monitoring transactions and detecting fraudulent activities. Columns may include transaction details, payment methods, and fraud indicators. (III) Product Search. E-commerce platforms employ tabular search and retrieval for product searches. Users can search for products based on various criteria, such as category, price range, and specifications, and the system retrieves matching products. (IV) tabular Transference. Transfer learning allows e-commerce businesses to leverage existing knowledge from one context to improve performance in related tasks, reducing the need for extensive training on new datasets. It can lead to more efficient model training, faster adaptation to new markets Real-world application relates to Supply Chain Optimization, Dynamic Pricing prediction and Customer Lifetime Value Prediction.

Energy Management Energy Management: Utilities and energy companies use tabular data to monitor energy consumption, analyze usage patterns, and optimize distribution networks for efficient energy management. Realworld application relates to peak demand prediction, Renewable Energy Source Prediction, Carbon Emission Data Retrieval

# 3 Tabular Data Modeling

In this section, we present an elaborated taxonomy of three aspects in tabular data learning: **Attributes Representation, Inter-Column Dependency Modeling, Side Learning Tasks**. We deliver a thorough analysis, delving into their primary challenges and typical strategies.

#### 3.1 Heterogeneous Attributes Encoding

Challenges Tabular learning aims to capture the intrinsic characteristics of each attribute (column). However, given that different tabular columns often originate from diverse sources and exhibit a high degree of heterogeneity, adopting a uniform encoding strategy becomes suboptimal. Current research predominantly focuses on three key dimensions within this domain:

Sub-Challenge 1: Diverse Data Formats and Distributions Heterogeneous tabular data frequently encompasses attribute formats including real-valued variables, categorical tags, textual descriptions, etc. Attributes within the same format may also display distinct distributions and ranges. This diversity necessitates the design of attribute-specific encoding strategies—for example, encoding real-valued variables using learned Gaussian models and mapping categorical tags to vector spaces. As a result, learning a homogeneous representation space for all attributes poses a challenge but is crucial for facilitating model training and designing effective loss functions.

**Sub-Challenge 2: Attributes Representation Spaces.** Data is inherently composed of elementary units representing the fundamental components that may not be further decomposed. Choices of elementary units involve trade-offs between representation ability, interpretability, and robustness. For instance, an image of a bird can be depicted as a collection of pixels or a composition of bird parts. Consequently, tabular representation learning also revolves around identifying the basic elements within tabular columns to devise effective representation strategies.

**Sub-Challenge 3: Incorporating Semantic Domain Knowledge.** The magnitude of a cell in isolation may lack meaningful interpretation without contextual information from the tabular header. For instance, the value "100" can convey distinct meanings depending on whether it is associated with 'PH' or 'Heart Rate'. Enhancing tabular representation, generalization, and robustness involves exploring, harnessing, and integrating rich contextual semantics from headers into tabular cells. This integration is crucial for ensuring the meaningful interpretation of tabular data in various contexts.

**Solution to Sub-Challenge 1.** Addressing the challenge of diverse data, it is proposed to conduct homogeneous learning and transform the raw mixed-type attributes into a unified space with continuous properties, which can also ease the gradient-based learning pipeline. A popular strategy is feature tokenizer, that converts each numerical and categorical column into a d-dimensional vector, e.g., a linear transformation [Huang  $et\ al.$ , 2020; Gorishniy  $et\ al.$ , 2021]. Another strategy involves a two-stage training process where the first stage aims to generate a more homogeneous representation of the data across dimensions, subsequently utilized by

the second stage. Typically, after the generative model, e.g., VAE has been effectively trained, the latent embeddings are extracted through the encoder and serve as the input for the subsequent stage.

Solution to Sub-Challenge 2. In addition to previous studies that focused on converting raw tabular data, especially categorical information, into a continuous space, the second group of research concentrates on the fundamental elements of tabular representation learning. Taking inspiration from the widely used one-hot encoding algorithm for categorical features discretizes numerical features into intervals, replacing original values with discrete descriptors. These descriptors can be represented using piecewise linear encoding or periodic activation functions. Another approach involves representing each cell as a new form of "column header is value." A concatenation of these cells is considered a fundamental representation of each row record.

Solution to Sub-Challenge 3. The key insight in this direction is that there is often abundant, auxiliary domain information that can further describing input feature semantic representations and interactions [Nguyen et al., 2019; Santos et al., 2022]. Research works in this direction aims to enhance semantic representation through the usage of external knowledge like knowledge graphs [Liu et al., 2023]. The learning procedure can be summarized as the following steps: (1) **Knowledge-graph construction** they construst an auxiliary KG to describe input feature, in which each input feature corresponds to a node in the auxiliary KG. (2) Node Embedding. Each input feature j associated to a learnable weight vector  $\theta_i \in \mathbb{R}^h$ , e.g., MLP, such that the weight vectors of all d features compose the weight matrix  $\theta \in \mathbb{R}^{d \times h}$ . (3) **Fea**ture interaction estimation. A trainable message-passing function could be learned to further update node embedding, based on the assumption that two input features which correspond to similar nodes in the KG should have similar weight vectors in the node embedding space [Ruiz et al., 2023].

Notably, tabular semantic representation is a prevalent topic in Natural Language Processing [Arik and Pfister, 2021], e.g., Tabular QA and Tabular semantic parsing. However, the majority of benchmark datasets in this domain enriches full-text descriptions, which falls outside the scope of this survey. Interested readers are directed to [] for a comprehensive exploration.

#### 3.2 Inter-Column Dependency Modeling

In tables, different columns may exhibit overlapping semantics and correlations, with decision-making often contingent upon latent factors behind certain interactions. For example, in a clinical setting, correlations exist behind patient demographic attributes such as gender, age and vulnerability to certain diseases. The patient's response to a specific medical diagnosis may also be intricately linked to their familial history of inherited diseases. Formulating data dependencies behind tabular columns improves representation learning and facilitates downstream decision-making, representing a central challenge.

In tables, different columns may exhibit overlapping semantics and correlations, with decision-making often contingent upon latent factors behind certain interactions. For example, in a clinical setting, correlations exist behind patient demographic attributes such as gender, age and vulnerability to certain diseases. The patient's response to a specific medical diagnosis may also be intricately linked to their familial history of inherited diseases. Formulating data dependencies behind tabular columns improves representation learning and facilitates downstream decision-making, representing a central challenges. Challenges In tabular data, different attributes work complementary to each other and intricate relationships often exist behind their observed values. Capturing these dependency structures is crucial for modeling the observed data and enhancing representation capability. Moreover, deep neural networks (DNNs) encounter challenges when learning from dense numerical tabular features due to the complexity of optimization hyperplanes in fully connected models, which increases the risk of converging to local optima [Fernández-Delgado et al., 2014]. Structure modeling can alleviate this challenge by uncovering critical factors and ease the learning burden. Commonly employed data structures include trees [Silva et al., 2020; Zhao et al., 2022a], graphs [Zhou et al., 2022; Zhao et al., 2024], and capsules [Chen et al., 2022] and logic rules [Ren et al., 2024]. For example, tree-based method shows its merit in iteratively picking the features with the largest statistical information gain [Chen and Guestrin, 2016]. In this survey, we explore recent advancements to capture relations across tabular columns in three streads:

**Sub-Challenge 1: Hierarchical Structure Modeling.** Hierarchical Structure defines an arrangement where abstract concepts are delineated as a function of less abstract ones. This hierarchical relation is prevalent in tables. For example, a table may have columns representing country, region, and city, creating a hierarchy where cities are nested within regions, and regions are nested within countries. Identifying, capturing, and modeling latent hierarchical structures constitutes a fundamental challenge in this context.

**Sub-Challenge 2: Interactive Structures Modeling.** In tables, different columns may exhibit overlapping semantics and correlations, with decision-making often contingent upon latent factors behind certain interactions. For example, in a clinical setting, correlations exist behind patient demographic attributes such as gender, age and vulnerability to certain diseases. The patient's response to a specific medical diagnosis may also be intricately linked to their familial history of inherited diseases. Formulating data dependencies behind tabular columns improves representation learning and facilitates downstream decision-making, representing a central challenge.

**Sub-Challenge 3: Latent Structure Discovery.** The challenges presented by Challenge 1 and Challenge 2 raise a fundamental question regarding the modeling of data dependency structures. However, such structure is usually latent and unknown, particularly in situations where domain-specific knowledge is absent. The discovery of such structures is essential for advancing structural modeling and serves as valuable evidence for root cause analysis, enhancing the explanatory capabilities of the model in a data-driven way.

Solution to Sub-Challenge 1: Depending on the target of specific structure, such methods can be divided

into three sub-directions. Sub-Solution 1a: tree structure modeling. DeepGBM [Ke et al., 2019] is a pioneering work that integrates the advantages of DNN and GBDT. DeepGBM comprises two distinct neural network components: CatNN, specialized in managing sparse categorical features, and GBDT2NN, tailored for distilling knowledge from GBDT to process dense numerical features. The NODE architecture [Popov et al., 2019] generalizes ensembles of oblivious decision trees, harnessing the advantages of both end-to-end gradient-based optimization and the efficacy of multi-layer hierarchical representation learning. GrowNet [Badirli et al., 2020] employs shallow neural networks as weak learners within a versatile gradient boosting framework. [Good et al., 2023] introduce an innovative framework that pioneers the alternation between sparse feature learning and differentiable decision tree construction. This approach aims to generate small, interpretable trees while maintaining high performance.

Sub-Solution 1 b: neural module-based structure learning. TABCAPs [Chen et al., 2022] utilize capsule networks to model feature-wise interactions. In the primary capsule layer, each sample undergoes encoding into multiple vectorial features using optimizable multivariate Gaussian kernels []. Subsequently, a successive iterative process of feature clustering is applied to attain higher-level semantics. TANGOS [Jeffares et al., 2022] utilizes a sparse and orthogonal regularization on the neural network, encouraging latent neurons to emphasize sparse, non-overlapping input features. This approach results in a collection of diverse and specialized latent units. Net-DNF [Katzir et al., 2020] presents an innovative framework characterized by an inherent inductive bias that yields models structured according to logical Boolean formulas in disjunctive normal form (DNF) over affine softthreshold decision terms. Moreover, Net-DNFs actively promote localized decision-making processes executed over discrete subsets of the input features.

Sub-Solution 1c: graph structure modeling, with a considerable body of research dedicated to this direction. These approaches typically construct a graph based on predefined node and edge categories before using (heterogeneous variants of) graph neural networks to capture a representation of structure. Among these approaches, graph neural networks (GNN) are commonly employed to model (i) Featurewise Interactions that Captures interactions and dependencies between individual features; (ii) Instance-wise Relations that models relationships between different instances or rows of the tabular data; and (iii) Feature-Instance correlations that address correlations between features and specific instances within the table. Feature-wise graph modeling aims to automatically estimate and represent relations among tabular features in the form of a learnable weighted graph [Zhou et al., 2022; Yan et al., 2023]. The learning procedure can be summarized into the following steps: (1) **Graph Con**struction. They represent each feature as a node and estimate pair-wise feature interactions as an adjacency matrix. (2) Graph Structure Learning. To estimate expressive feature relations and learn a reliable graph structure, prior works primarily focus on learning the adjacency matrix, considering three key factors: 1) node semantic meaning; 2) implicit feature relations [Zhao et al., 2022b]; 3) homogeneous assumption in GNN that heterogeneous tabular property violates the assumption of GNNs that connected nodes should exhibit similar patterns [Zhu et al., 2020]. [Yan et al., 2023] generates data-adaptive edge weights by performing node semantic matching. [Zhou et al., 2022] both consider node semantic matching and high-order interaction and achieve this by an ensemble graph that consists of an adaptive probability adjacency matrix with self attention and a static graph which calculate relation topology score through node semantic embeddings. Recently, [Wan et al., 2023] find that features from multiple fields in tabular data exhibit different patterns and it is difficult for all of them to be matched with one single graph. To address this, they first perform a hierarchically clustering method to separates initial tabular data into several parts with minimal correlation and then derived graph structures from decorrelated clusters that have minimal feature correlations. (3) Graph Sparfication. To remove redundant feature relations and improve training efficiency without a constructed dense graph, they selectively collects salient features for the final prediction []. [Yan et al., 2023] design a global readout node to selectively collect salient features from each layer. [Zhou et al., 2022] employ the reinforcement learning method to strengthen the key feature interaction connections. Instance-wise graph modeling treats each instance as a node in a graph and cast tabular classification as node classification problem [Liao and Li, 2023]. The third line of feature-instance graph modeling aim to model instance-feature associates and perform message Propagation to enhance the target data instance representations. They construct a data-feature bipartite graph with data instance nodes and feature value nodes and then perform message passing on hypergraph neural network [Du et al., 2022].

Solution to Sub-Challenge 2. Interactive Structures Modeling formulates data dependency in a sequential decisionmaking framework, where the high-level concept depends on the local decisions made at every step. Sub-solution 2a comprises transformer-based methods that leverage sequential attention mechanisms to determine which features to consider at each decision step. This approach enhances interpretability and improves learning efficiency by allocating learning capacity to the most salient features. For instance, TabNet [Arik and Pfister, 2021] designs a sequential attention mechanism for both feature selection and reasoning. SAINT [Somepalli et al., 2022] introduces the Self-Attention and Intersample Attention Transformer, allowing attention mechanisms over both rows and columns. NAN [Luo et al., 2020] introduces a Network On Network (NON) architecture tailored for tabular data classification. NON comprises three integral components: a field-wise network at the base, designed to capture intra-field information; an across-field network in the middle, dynamically selecting suitable operations based on data characteristics; and an operation fusion network at the summit, facilitating deep fusion of outputs from the chosen operations.

**Sub-Solution 2b** investigated using Differentiable Decision Tree (DDTs) for reinforcement learning tasks, with the goal to combine the flexibility of neural networks with the interpretable structure of decision trees. Existing works

focus on policy learning using DDTs [Silva et al., 2020; Tambwekar et al., 2023; Pace et al., 2022], on the contrary, [Kalra and Brown, 2023] aims to learn interpretable reward functions using DDTs.

**Solution to Sub-Challenge 3**. Latent Structure Discovery aims to learn latent dependencies in a fully data-driven manner. Defining the structure of tabular data can be challenging, especially without domain knowledge. As an alternative approach, recent efforts have explored the use of differentiable strategies to autonomously search for the tabular structure, as seen in AGEBO-Tabular [Egele *et al.*, 2021] and Tab-NAS [Yang *et al.*, 2022]. Another direction involves leveraging generative models for discovering latent structures, which will be elaborated upon in Section 3.3.

## 3.3 Specialized Learning Tasks on Tabular Data

In addition to strategies that enhance data representations by attribute-specific encoding and relation modeling, a myriad of works focuses on specific tasks and learning goals on tabular data. These tasks are critical for the broader application of tables, and may provide additional learning signals to improve tabular representations. In this section, we introduce two representative side learning tasks, generative modeling of tabular data, and knowledge transfer across tables in similar domains.

#### **Tabular Data Generation Tasks**

Tabular data is a prevalent data format in various industrial sectors, including finance and electronic health records. However, the limited privacy considerations constrain the feasibility of releasing such data. Moreover, the collection of such data involves intricate procedures, such as unbalanced data nature, ensuring participant willingness and synchronized updates. Consequently, collected tabular datasets often contain missing values. The generation of authentic tabular data is crucial, particularly in the contexts of tabular data imputation and synthetic data cre-Existing works primarily devise model architectures based on popular generative models, including GAN [Goodfellow et al., 2020], VAE [Kingma, 2013], and the diffusion model [Wijmans and Baker, 1995]. These efforts aim to tackle three core challenges prevalent in tabular data generation: 1) Input Heterogeneity; 2) Sampling Quality; 3) Latent Structure Modeling.

As a pionering work in the realm of Generative Adversarial Networks (GANs), medGAN, as described by [Choi et al., 2017], integrates both an auto-encoder and a GAN to effectively generate diverse sets of continuous and/or binary medical data. Subsequently, TableGAN [Park et al., 2018] employs a Convolutional Neural Network (CNN) for feature processing. To address these complexities in modeling numerical features, CTGAN [Zhao et al., 2021] enhances its training procedure by incorporating modespecific normalization and mitigates data imbalance through the implementation of a conditional generator.

The second group leverage Variational Autoencoder (VAE) methodologies, employing the well-known 'encode-sampling-decode' pipeline. VAEM [Ma *et al.*, 2020] undergoes a two-stage training process. A Variational Autoencoder (VAE) is employed to establish a more homogeneous latent

representation across dimensions in the first stage. Subsequently, this continuously extracted latent representation is extracted and further utilized in a second VAE for the purpose of dependency modeling. Building upon the principles introduced by VAEM, TABSYN [Zhang et al., 2023] incorporates the extracted latent representation into a diffusion model while HH-VAE [Peis et al., 2022] introduces Hamiltonian Monte Carlo to improved approximate inference quality. Contrary to the previously mentioned approaches, GOGGLE [Liu et al., 2022] focuses on learning a graph structure within the latent space.

The third group are constructed based on a diffusion model [], which approximates the target distribution by examining the endpoint of a Markov chain. This chain originates from a specified parametric distribution, commonly chosen as a standard Gaussian distribution. CoDi [Lee et al., 2023] and TABDDPM [Kotelnikov et al., 2023] are two concurrent works that leverage Gaussian diffusion models and Multinomial diffusion models to formulate numerical and categorical data seperatelly. Current works mainly borrow a similar idea on application-driven tasks, e.g., finance domain [Sattarov et al., 2023]. Differently, STaSy [Kim et al., 2022a] proposes a self-paced learning technique and a fine-tuning strategy, which further increases the sampling quality and diversity by stabilizing the denoising score matching training. SOS [Kim et al., 2022b] oversamples minor classes since imbalanced classes frequently lead to suboptimal training outcomes.

#### **Tabular Data Imputation Tasks**

Another line of works address the missing value problems and focus on data imputation. Predictive approaches to missing data imputation can be categorized in two families [Telyatnikov and Scardapane, 2023]: (i) imputing missing data through estimating statistics using the entire dataset [Lakshminarayan et al., 1996; Nazabal et al., 2020]; (ii) inferring the missing components employing similar data points to the one having missing values, e.g, KNN-based method [Acuna and Rodriguez, 2004]. To both model feature interaction in a global manner and leverage similar sample information, recent works [Spinelli et al., 2020] [Telyatnikov and Scardapane, 2023] explored the assumption of endowing tabular data with a graph topology and then exploiting the message mechanism to impute missing value. The difficulty and challenges here remain in the definition of a suitable distance metric to compute graph connectivity beforehand and design a customized procedure to sparsify the graph [Telyatnikov and Scardapane, 2023].

#### **Pretraining on Tabular Data**

Challenges Tabular transference aims to transfer knowledge between tables. Tabular data exhibit variations in both the number and types of columns (termed as variable-column), posing a challenge for tabular deep learning models to transfer knowledge effectively from one table to another. Besides, in contrast to image and text modalities, tabular data are highly domain-specific and often lack extensive, high-quality datasets. These three challenges can result in poor generalization abilities across diverse tabular datasets. Based on the

transfering difficulty and data distribution shift types, the core research questions can be classified into the following streads:

**Sub-Challenge 1: Covariant Shift.** Covariate shift is a common scenario in industrial practice, signifying a shift in the marginal distribution of features while the decision boundary of the model, denoted as p(y|x), remains unaltered. For instance, a patient may undergo multiple visits to a clinical institution, leading to a shifted feature representation. However, the underlying mechanism p(y|x) remains stable.

Sub-Challenge 2: Distribution Shift with Varied Columns. In industrial practice, feature design constitutes a critical step where scientists and engineers may introduce new features while removing redundant or unnecessary features. Given the substantial overlap in features, retraining the entire model is time-consuming and results in inefficiencies in labor. However, formulating an effective transfer learning paradigm that incorporates new feature information while discarding removed feature knowledge is a non-trivial task.

**Sub-Challenge 3: Distribution Shift across Domains** Vision and text models exhibit adaptability to a diverse array of tasks. This adaptability is attributed to the shared general representations present in both sentences and images, which shows task-agnostic properties [Farahani *et al.*, 2021]. However, in the context of heterogeneous data [Ren *et al.*, 2022a], a pertinent question arises: is there shared knowledge across tables, considering that two distinct tables can possess entirely different column numbers and associated semantic meanings?

Solution to Sub-Challenge 1 is within-table pretraining that only covariate shift occurs. The primary objective is to devise a self-supervised loss and create a corrupted or augmented rendition of the initial tabular data. In the case of VIME [Yoon et al., 2020], a mask matrix  $\mathcal{M}$  is applied to the initial tabular data  $\mathcal{X}$  to generate a corrupted version  $\mathcal{X}$  as input. The model then undertakes tabular data reconstruction and mask vector estimation, constituting two self-supervised loss components. Conversely, the subsequent work, SCARF [Bahri et al., 2021], generates a corrupted version by replacing each feature with a random draw from its empirical marginal distribution. Additionally, it introduces a contrastive loss. In contrast to these approaches, SubTab [Ucar et al., 2021] posits that reconstructing the data from a subset of its features, rather than from its corrupted version in an autoencoder setting, can better capture the underlying latent representation.

**Solution to Sub-Challenge 2** is across-table pretraining that formulates a versatile model capable of accommodating variable-column tables. One direction is to convert tabular data (cells in columns) into a sequence of semantically encoded tokens, e.g., TransTab [Wang and Sun, 2022], MED ITAB [Wang et al., 2023], TABRET[Onishi et al., 2023], UniTabE [Yang et al., 2023]. Another line of works aim to propose a pseudo-feature method for aligning the upstream and downstream feature sets in heterogeneous data, e.g., [Levin et al., 2022].

**Sub-Solution 3a to Sub-Challenge 3:** is Cross-domain pretraining that goes beyond the limitations imposed by variable-column structures and domain-specific constraints. XTab [Zhu *et al.*, 2023], as a pioneering work, engages in

Representation	Homogeneous Representation	[Huang et al., 2020], [Gorishniy et al., 2021]
	Element Representation	
	Semantic Representation	
Dependency modeling	Hierarchical structure modeling	DeepGBM [Ke et al., 2019], [Popov et al., 2019], GrowNet [Badirli et al., 2020], [Good et al., 2023], TABCAPs [Chen et al., 2022], TANGOS [Jeffares et al., 2022], Net-DNF [Katzir et al., 2020], [Zhou et al., 2022], [Yan et al., 2023], [Wan et al., 2023], [Liao and Li, 2023], [Du et al., 2022]
	Interactive Structures Modeling	TabNet [Arik and Pfister, 2021], SAINT [Somepalli <i>et al.</i> , 2022], NAN [Luo <i>et al.</i> , 2020], [Silva <i>et al.</i> , 2020], [Tambwekar <i>et al.</i> , 2023], [Pace <i>et al.</i> , 2022], [Kalra and Brown, 2023]
	Latent Structure Discovery	AGEBO-Tabular [Egele et al., 2021], TabNAS [Yang et al., 2022]
Generation	Generative Adversarial Networks	medGAN [Choi <i>et al.</i> , 2017], TableGAN [Park <i>et al.</i> , 2018], CTGAN [Zhao <i>et al.</i> , 2021]
	Variational Autoencoder	VAEM [Ma et al., 2020], TABSYN [Zhang et al., 2023], HH-VAE [Peis et al., 2022], GOGGLE [Liu et al., 2022]
	Diffusion Model	CoDi [Lee <i>et al.</i> , 2023], TABDDPM [Kotelnikov <i>et al.</i> , 2023], [Sattarov <i>et al.</i> , 2023], STaSy [Kim <i>et al.</i> , 2022a], SOS [Kim <i>et al.</i> , 2022b]
Transference	within table pretraining	VIME [Yoon <i>et al.</i> , 2020], SCARF [Bahri <i>et al.</i> , 2021], SubTab [Ucar <i>et al.</i> , 2021]
	across table pretraining	TransTab [Wang and Sun, 2022], MED ITAB [Wang et al., 2023], TABRET[Onishi et al., 2023], UniTabE [Yang et al., 2023], [Levin et al., 2022]
	across domain pretraining	XTab [Zhu <i>et al.</i> , 2023]
	transference with Large Language Model	TabLLM [Hegselmann et al., 2023], Anypredict [Wang et al., 2023], GReaT [Borisov et al., 2022b], Graphcare [Jiang et al., 2023], CHAIN-OF-TABLE [Wang et al., 2024b]

Table 1: An example table.

pretraining on a diverse set of over 150 tables collected from finance, education, and medical domains. To address column variate challenges, Xtab [Zhu *et al.*, 2023] utilize independent featurizers and use federated learning to pretrain the shared component. Notably, XTab doesn't strive to learn a universal tokenizer applicable to all tables. Instead, its goal is to acquire a weight initialization that exhibits generalizability across various downstream tasks.

**Sub-Solution 3b to Sub-Challenge 3.** Currently, transference with LLMs emerges as a new direction for across-domain transference. TabLLM [Hegselmann *et al.*, 2023], Anypredict [Wang *et al.*, 2023], GReaT [Borisov *et al.*, 2022b] first serialize the feature names and values into a natural language string. This string is then combined with a task-specific prompt for model fine-tuning. Recently, Graphcare [Jiang *et al.*, 2023] extracts healthcare-related knowledge graph (KG) from LLMs. The extracted KG is expected to embed generalizable representation and can be used to improve EHR-based predictions.

# 4 Conclusion and Future Directions

#### 4.1 Representation

**&** Benchmarks. Tabular data stands as a prevalent data format in industrial settings; however, its accessibility is fre-

quently constrained due to concerns surrounding data privacy, particularly in domains such as clinical and finance. The release and synthesis of high-quality tabular data would significantly enhance algorithmic design and contribute substantively to the advancement of representation learning.

- ♣ Theoretical Analysis. Existing works achieve promising empirical performance on tabular data modeling problems, the theoretical analysis remains an open problem. We believe that rigorous analyses can provide in-depth insights and inspire the development of new tabular-based methods. Here we propose several research questions: (1) How to formulate and analyze the influence of missing-value problem in tabular representation? (2) Different tabular columns generally exhibit a diversity of distribution statistics, e.g., mean, variance. How to evaluate and quantify the significance of such variability in representation learning and loss design?
- ♣ Empirical Analysis A pioneering work in [] demonstrates that shallow layer in DNN shares generalizable features, e.g., image edge, texture, which builds empirical foundations for transfer learning in image domain. However, empirical validation regarding the transferability and generalization of knowledge within the framework of tabular representations remains largely unexplored. Further exploration is required to comprehensively investigate and validate these aspects within the tabular domain.

- ♣ Element Representation. Existing works mainly consider tabular heterogeneous nature and transform mixed type feature into a unified continuous space. One pioneering work, [], first split numerical features into bins and and find that bin-based embedding could large improve performance. The obtention of element units, especially from the perspective of available external knowledge, is still an under-explored problem in the tabular data domain.
- **&** Semantic Representation. Given our primary focus on tabular data featuring numerical and categorical features, there are limited transferable representations within these values. However, tabular data is inherently domain-specific, implying that it should exhibit rich domain knowledge. Exploring the usage and interaction with large language models represents a promising direction. This approach would offer auxiliary information, enhancing both the semantic and generalizable representation of tabular data.

# 4.2 Dependency Modeling

- ♣ Temporal Dependency Modeling Current studies predominantly address static scenarios wherein tabular features remain constant. Real-world datasets frequently manifest temporal properties, such as sequential Electronic Health Record (EHR) tabular data and sequential financial data. Exploring methodologies for learning dynamic dependencies within temporal tabular data represents a practical and promising direction.
- ♣ Dependency Modeling with Auxiliary Knowledge Due to privacy concerns and labor costs, prior research endeavors have explored and modeled tabular data structures without incorporating external knowledge. With the advent of Large Language Models (LLMs) and RAG tools [Zhang et al., 2024], the prospect of constructing a trustworthy database and knowledge graph to enhance the quality of dependency discovery and the modeling of dependency structures emerges as an intriguing and promising direction.
- ♣ Modeling Data Dependency with Missing Features and Labels In practical settings, tabular data often encompasses instances with missing features or labels. For instance, within Electronic Health Record (EHR) data, certain patient records may exhibit absence of values pertaining to specific medical tests, diagnoses, or other health-related features. The issue of missingness has been underexplored in existing literature. One important direction would be develop robust dependency models that can effectively handle such scenarios. This involves designing algorithms and approaches that can infer dependencies, patterns, or correlations in the presence of missing data, ensuring that the learned dependencies remain applicable and informative even when confronted with incomplete information.

#### 4.3 Generation

♣ Sampling quality Current research primarily relies on generative models for the purpose of synthesizing artificial tabular data. However, realistic tabular data are imbalanced, missingness and heterogeneous. A classical generative model exhibits a tendency to emphasize the majority class while overlooking its minority components, resulting in substantial biases affecting the quality of generated data. Further-

- more, the presence of missing data introduces significant uncertainty during the generation process. Addressing this challenge involves the incorporation of probabilistic models, representing a crucial direction for improvement. Besides, Recent work [] has observed that the heterogeneous nature of tabular data causes mode collapse in the latent space when employing VAE-based methods. Considering the heterogeneous property in the latent space and improve sampling quality is an intriguing and important research problem.
- ♣ Structure Modeling. Contemporary research predominantly explores the utilization of advanced generative models for synthesizing tabular data, e.g, from GAN to diffusion models. However, the incorporation of dependency modeling in the generation process is frequently overlooked in recent works. Leveraging the inherent strengths of data structure modeling, e.g., integrate and adapt tree structure into diffusion model stands as an interesting and challenging research question.

#### 4.4 Transference

- Adaptation without Accessing to Source Data In an industrial setting, retraining models is expensive and time-consuming [Ren et al., 2022b]. A more realistic setting considering covariate shift is tabular test time adaptation, where only pretrained model and target tabular feature are available. The development of a tailored loss function and adaptation model designed specifically for tabular data holds the potential to yield substantial cost savings for industry companies.
- Adaptation with Continual Distribution Shift. Current research predominantly focuses on the static nature of tabular data, which does not accurately reflect the dynamic nature inherent in realistic finance or clinical data. The development of a robust and generalizable model capable of handling the evolving distribution shift, e.g., open feature sets, in tabular data would capture the attention of the industry.
- Adaptation with Large Language Models. Given that tabular data is inherently domain-specific and may encompass diverse columns, previous works use column imputation strategies to ensure alignment across columns. However, the imputed data often contain significant noise. Exploring the utilization and integration of LLMs for column information imputation emerges as an intriguing and promising direction

### References

- [Acuna and Rodriguez, 2004] Edgar Acuna and Caroline Rodriguez. The treatment of missing values and its effect on classifier accuracy. In Classification, Clustering, and Data Mining Applications: Proceedings of the Meeting of the International Federation of Classification Societies (IFCS), Illinois Institute of Technology, Chicago, 15–18 July 2004, pages 639–647. Springer, 2004.
- [Arik and Pfister, 2021] Sercan Ö Arik and Tomas Pfister. Tabnet: Attentive interpretable tabular learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 6679–6687, 2021.
- [Badirli et al., 2020] Sarkhan Badirli, Xuanqing Liu, Zhengming Xing, Avradeep Bhowmik, Khoa Doan, and

- Sathiya S Keerthi. Gradient boosting neural networks: Grownet. *arXiv preprint arXiv:2002.07971*, 2020.
- [Bahri et al., 2021] Dara Bahri, Heinrich Jiang, Yi Tay, and Donald Metzler. Scarf: Self-supervised contrastive learning using random feature corruption. In *International Conference on Learning Representations*, 2021.
- [Borisov et al., 2022a] Vadim Borisov, Tobias Leemann, Kathrin Seßler, Johannes Haug, Martin Pawelczyk, and Gjergji Kasneci. Deep neural networks and tabular data: A survey. IEEE transactions on neural networks and learning systems, 2022.
- [Borisov *et al.*, 2022b] Vadim Borisov, Kathrin Seßler, Tobias Leemann, Martin Pawelczyk, and Gjergji Kasneci. Language models are realistic tabular data generators. *arXiv preprint arXiv:2210.06280*, 2022.
- [Chen and Guestrin, 2016] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings* of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pages 785–794, 2016.
- [Chen et al., 2022] Jintai Chen, KuanLun Liao, Yanwen Fang, Danny Chen, and Jian Wu. Tabcaps: A capsule neural network for tabular data classification with bow routing. In *The Eleventh International Conference on Learning Representations*, 2022.
- [Choi et al., 2017] Edward Choi, Siddharth Biswal, Bradley Malin, Jon Duke, Walter F Stewart, and Jimeng Sun. Generating multi-label discrete patient records using generative adversarial networks. In *Machine learning for health-care conference*, pages 286–305. PMLR, 2017.
- [Du et al., 2022] Kounianhua Du, Weinan Zhang, Ruiwen Zhou, Yangkun Wang, Xilong Zhao, Jiarui Jin, Quan Gan, Zheng Zhang, and David P Wipf. Learning enhanced representation for tabular data via neighborhood propagation. Advances in Neural Information Processing Systems, 35:16373–16384, 2022.
- [Egele et al., 2021] Romain Egele, Prasanna Balaprakash, Isabelle Guyon, Venkatram Vishwanath, Fangfang Xia, Rick Stevens, and Zhengying Liu. Agebo-tabular: joint neural architecture and hyperparameter search with autotuned data-parallel training for tabular data. In Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, pages 1–14, 2021.
- [Farahani et al., 2021] Abolfazl Farahani, Sahar Voghoei, Khaled Rasheed, and Hamid R Arabnia. A brief review of domain adaptation. Advances in data science and information engineering: proceedings from ICDATA 2020 and IKE 2020, pages 877–894, 2021.
- [Fernández-Delgado *et al.*, 2014] Manuel Fernández-Delgado, Eva Cernadas, Senén Barro, and Dinani Amorim. Do we need hundreds of classifiers to solve real world classification problems? *The journal of machine learning research*, 15(1):3133–3181, 2014.
- [Good et al., 2023] Jack Henry Good, Torin Kovach, Kyle Miller, and Artur Dubrawski. Feature learning for inter-

- pretable, performant decision trees. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.
- [Goodfellow *et al.*, 2020] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.
- [Gorishniy et al., 2021] Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting deep learning models for tabular data. Advances in Neural Information Processing Systems, 34:18932–18943, 2021.
- [Hegselmann et al., 2023] Stefan Hegselmann, Alejandro Buendia, Hunter Lang, Monica Agrawal, Xiaoyi Jiang, and David Sontag. Tabllm: Few-shot classification of tabular data with large language models. In *International Conference on Artificial Intelligence and Statistics*, pages 5549–5581. PMLR, 2023.
- [Huang *et al.*, 2020] Xin Huang, Ashish Khetan, Milan Cvitkovic, and Zohar Karnin. Tabtransformer: Tabular data modeling using contextual embeddings. *arXiv* preprint arXiv:2012.06678, 2020.
- [Jeffares et al., 2022] Alan Jeffares, Tennison Liu, Jonathan Crabbé, Fergus Imrie, and Mihaela van der Schaar. Tangos: Regularizing tabular neural networks through gradient orthogonalization and specialization. In *The Eleventh International Conference on Learning Representations*, 2022.
- [Jiang et al., 2023] Pengcheng Jiang, Cao Xiao, Adam Cross, and Jimeng Sun. Graphcare: Enhancing health-care predictions with open-world personalized knowledge graphs. arXiv preprint arXiv:2305.12788, 2023.
- [Kalra and Brown, 2023] Akansha Kalra and Daniel S Brown. Can differentiable decision trees learn interpretable reward functions? arXiv preprint arXiv:2306.13004, 2023.
- [Katzir *et al.*, 2020] Liran Katzir, Gal Elidan, and Ran El-Yaniv. Net-dnf: Effective deep modeling of tabular data. In *International conference on learning representations*, 2020.
- [Ke et al., 2019] Guolin Ke, Zhenhui Xu, Jia Zhang, Jiang Bian, and Tie-Yan Liu. Deepgbm: A deep learning framework distilled by gbdt for online prediction tasks. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pages 384– 394, 2019.
- [Kim et al., 2022a] Jayoung Kim, Chaejeong Lee, and Noseong Park. Stasy: Score-based tabular data synthesis. In *The Eleventh International Conference on Learning Representations*, 2022.
- [Kim et al., 2022b] Jayoung Kim, Chaejeong Lee, Yehjin Shin, Sewon Park, Minjung Kim, Noseong Park, and Jihoon Cho. Sos: Score-based oversampling for tabular data. In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pages 762–772, 2022.

- [Kingma, 2013] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
- [Kotelnikov *et al.*, 2023] Akim Kotelnikov, Dmitry Baranchuk, Ivan Rubachev, and Artem Babenko. Tabddpm: Modelling tabular data with diffusion models. In *International Conference on Machine Learning*, pages 17564–17579. PMLR, 2023.
- [Lakshminarayan et al., 1996] Kamakshi Lakshminarayan, Steven A Harp, Robert P Goldman, Tariq Samad, et al. Imputation of missing data using machine learning techniques. In KDD, volume 96, 1996.
- [Lee *et al.*, 2023] Chaejeong Lee, Jayoung Kim, and Noseong Park. Codi: Co-evolving contrastive diffusion models for mixed-type tabular synthesis. *arXiv preprint arXiv:2304.12654*, 2023.
- [Levin et al., 2022] Roman Levin, Valeriia Cherepanova, Avi Schwarzschild, Arpit Bansal, C Bayan Bruss, Tom Goldstein, Andrew Gordon Wilson, and Micah Goldblum. Transfer learning with deep tabular models. arXiv preprint arXiv:2206.15306, 2022.
- [Li et al., 2019] Ziyan Li, Jianjiang Feng, Zishun Feng, Yunqiang An, Yang Gao, Bin Lu, and Jie Zhou. Lumen segmentation of aortic dissection with cascaded convolutional network. In Statistical Atlases and Computational Models of the Heart. Atrial Segmentation and LV Quantification Challenges: 9th International Workshop, STACOM 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers 9, pages 122–130. Springer, 2019.
- [Liang et al., 2024] Junjie Liang, Weijieying Ren, Hanifi Sahar, and Vasant Honavar. Inducing clusters deep kernel gaussian process for longitudinal data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 13736–13743, 2024.
- [Liao and Li, 2023] Jay Chiehen Liao and Cheng-Te Li. Tabgsl: Graph structure learning for tabular data prediction. *arXiv preprint arXiv:2305.15843*, 2023.
- [Liu et al., 2017] Honghui Liu, Jianjiang Feng, Zishun Feng, Jiwen Lu, and Jie Zhou. Left atrium segmentation in ct volumes with fully convolutional networks. In Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MIC-CAI 2017, Québec City, QC, Canada, September 14, Proceedings 3, pages 39–46. Springer, 2017.
- [Liu et al., 2022] Tennison Liu, Zhaozhi Qian, Jeroen Berrevoets, and Mihaela van der Schaar. Goggle: Generative modelling for tabular data by learning relational structure. In *The Eleventh International Conference on Learning Representations*, 2022.
- [Liu et al., 2023] Jixiong Liu, Yoan Chabot, Raphaël Troncy, Viet-Phi Huynh, Thomas Labbé, and Pierre Monnin. From tabular data to knowledge graphs: A survey of semantic table interpretation tasks and methods. *Journal of Web Semantics*, 76:100761, 2023.

- [Luo et al., 2020] Yuanfei Luo, Hao Zhou, Wei-Wei Tu, Yuqiang Chen, Wenyuan Dai, and Qiang Yang. Network on network for tabular data classification in real-world applications. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2317–2326, 2020.
- [Ma et al., 2020] Chao Ma, Sebastian Tschiatschek, Richard Turner, José Miguel Hernández-Lobato, and Cheng Zhang. Vaem: a deep generative model for heterogeneous mixed type data. Advances in Neural Information Processing Systems, 33:11237–11247, 2020.
- [Ma et al., 2022a] Haixu Ma, Yufeng Liu, and Guorong Wu. Elucidating multi-stage progression of neuro-degeneration process in alzheimer's disease. Alzheimer's & Dementia, 18:e068774, 2022.
- [Ma et al., 2022b] Haixu Ma, Donglin Zeng, and Yufeng Liu. Learning individualized treatment rules with many treatments: A supervised clustering approach using adaptive fusion. Advances in Neural Information Processing Systems, 35:15956–15969, 2022.
- [Ma et al., 2023] Haixu Ma, Donglin Zeng, and Yufeng Liu. Learning optimal group-structured individualized treatment rules with many treatments. *Journal of Machine Learning Research*, 24(102):1–48, 2023.
- [Mackinnon, 2010] A Mackinnon. The use and reporting of multiple imputation in medical research—a review. *Journal of internal medicine*, 268(6):586–593, 2010.
- [Nazabal *et al.*, 2020] Alfredo Nazabal, Pablo M Olmos, Zoubin Ghahramani, and Isabel Valera. Handling incomplete heterogeneous data using vaes. *Pattern Recognition*, 107:107501, 2020.
- [Nguyen *et al.*, 2019] Phuc Nguyen, Natthawut Kertkeidkachorn, Ryutaro Ichise, and Hideaki Takeda. Mtab: Matching tabular data to knowledge graph using probability models. *arXiv preprint arXiv:1910.00246*, 2019.
- [Onishi *et al.*, 2023] Soma Onishi, Kenta Oono, and Kohei Hayashi. Tabret: Pre-training transformer-based tabular models for unseen columns. *arXiv preprint arXiv:2303.15747*, 2023.
- [Pace *et al.*, 2022] Alizée Pace, Alex J Chan, and Mihaela van der Schaar. Poetree: Interpretable policy learning with adaptive decision trees. *arXiv preprint arXiv:2203.08057*, 2022.
- [Park et al., 2018] Noseong Park, Mahmoud Mohammadi, Kshitij Gorde, Sushil Jajodia, Hongkyu Park, and Youngmin Kim. Data synthesis based on generative adversarial networks. arXiv preprint arXiv:1806.03384, 2018.
- [Peis et al., 2022] Ignacio Peis, Chao Ma, and José Miguel Hernández-Lobato. Missing data imputation and acquisition with deep hierarchical models and hamiltonian monte carlo. Advances in Neural Information Processing Systems, 35:35839–35851, 2022.
- [Popov et al., 2019] Sergei Popov, Stanislav Morozov, and Artem Babenko. Neural oblivious decision ensembles for

- deep learning on tabular data. In *International Conference* on Learning Representations, 2019.
- [Qin et al., 2023] Wei Qin, Zetong Chen, Lei Wang, Yunshi Lan, Weijieying Ren, and Richang Hong. Read, diagnose and chat: Towards explainable and interactive llms-augmented depression detection in social media. arXiv preprint arXiv:2305.05138, 2023.
- [Ren and Honavar, 2024] Weijieying Ren and Vasant G Honavar. Esacl: An efficient continual learning algorithm. In *Proceedings of the 2024 SIAM International Conference on Data Mining (SDM)*, pages 163–171. SIAM, 2024.
- [Ren et al., 2022a] Weijieying Ren, Lei Wang, Kunpeng Liu, Ruocheng Guo, Lim Ee Peng, and Yanjie Fu. Mitigating popularity bias in recommendation with unbalanced interactions: A gradient perspective. In 2022 IEEE International Conference on Data Mining (ICDM), pages 438– 447. IEEE, 2022.
- [Ren et al., 2022b] Weijieying Ren, Pengyang Wang, Xi-aolin Li, Charles E Hughes, and Yanjie Fu. Semi-supervised drifted stream learning with short lookback. In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pages 1504–1513, 2022.
- [Ren et al., 2024] Weijieying Ren, Xiaoting Li, Huiyuan Chen, Vineeth Rakesh, Zhuoyi Wang, Mahashweta Das, and Vasant G Honavar. Tablog: Test-time adaptation for tabular data using logic rules. In Forty-first International Conference on Machine Learning, 2024.
- [Ruan et al., 2024] Yucheng Ruan, Xiang Lan, Jingying Ma, Yizhi Dong, Kai He, and Mengling Feng. Language modeling on tabular data: A survey of foundations, techniques and evolution. arXiv preprint arXiv:2408.10548, 2024.
- [Ruiz et al., 2023] Camilo Ruiz, Hongyu Ren, Kexin Huang, and Jure Leskovec. High dimensional, tabular deep learning with an auxiliary knowledge graph. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.
- [Sahakyan *et al.*, 2021] Maria Sahakyan, Zeyar Aung, and Talal Rahwan. Explainable artificial intelligence for tabular data: A survey. *IEEE access*, 9:135392–135422, 2021.
- [Santos et al., 2022] Alberto Santos, Ana R Colaço, Annelaura B Nielsen, Lili Niu, Maximilian Strauss, Philipp E Geyer, Fabian Coscia, Nicolai J Wewer Albrechtsen, Filip Mundt, Lars Juhl Jensen, et al. A knowledge graph to interpret clinical proteomics data. Nature biotechnology, 40(5):692–702, 2022.
- [Sattarov et al., 2023] Timur Sattarov, Marco Schreyer, and Damian Borth. Findiff: Diffusion models for financial tabular data generation. In *Proceedings of the Fourth ACM International Conference on AI in Finance*, pages 64–72, 2023.
- [Sauber-Cole and Khoshgoftaar, 2022] Rick Sauber-Cole and Taghi M Khoshgoftaar. The use of generative adversarial networks to alleviate class imbalance in tabular data: a survey. *Journal of Big Data*, 9(1):98, 2022.

- [Silva et al., 2020] Andrew Silva, Matthew Gombolay, Taylor Killian, Ivan Jimenez, and Sung-Hyun Son. Optimization methods for interpretable differentiable decision trees applied to reinforcement learning. In *International conference on artificial intelligence and statistics*, pages 1855–1865. PMLR, 2020.
- [Somepalli et al., 2022] Gowthami Somepalli, Avi Schwarzschild, Micah Goldblum, C Bayan Bruss, and Tom Goldstein. Saint: Improved neural networks for tabular data via row attention and contrastive pre-training. In NeurIPS 2022 First Table Representation Workshop, 2022.
- [Spinelli *et al.*, 2020] Indro Spinelli, Simone Scardapane, and Aurelio Uncini. Missing data imputation with adversarially-trained graph convolutional networks. *Neural Networks*, 129:249–260, 2020.
- [Tambwekar et al., 2023] Pradyumna Tambwekar, Andrew Silva, Nakul Gopalan, and Matthew Gombolay. Natural language specification of reinforcement learning policies through differentiable decision trees. *IEEE Robotics and Automation Letters*, 2023.
- [Telyatnikov and Scardapane, 2023] Lev Telyatnikov and Simone Scardapane. Egg-gae: scalable graph neural networks for tabular data imputation. In *International Conference on Artificial Intelligence and Statistics*, pages 2661–2676. PMLR, 2023.
- [Ucar et al., 2021] Talip Ucar, Ehsan Hajiramezanali, and Lindsay Edwards. Subtab: Subsetting features of tabular data for self-supervised representation learning. Advances in Neural Information Processing Systems, 34:18853–18865, 2021.
- [Wan et al., 2023] Junhong Wan, Yao Fu, Junlan Yu, Weihao Jiang, Shiliang Pu, and Ruiheng Yang. Graphfade: Field-aware decorrelation neural network for graphs with tabular features. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 2502–2511, 2023.
- [Wang and Sun, 2022] Zifeng Wang and Jimeng Sun. Transtab: Learning transferable tabular transformers across tables. *Advances in Neural Information Processing Systems*, 35:2902–2915, 2022.
- [Wang et al., 2023] Zifeng Wang, Chufan Gao, Cao Xiao, and Jimeng Sun. Anypredict: Foundation model for tabular prediction. arXiv preprint arXiv:2305.12081, 2023.
- [Wang et al., 2024a] Wei-Yao Wang, Wei-Wei Du, Derek Xu, Wei Wang, and Wen-Chih Peng. A survey on selfsupervised learning for non-sequential tabular data. arXiv preprint arXiv:2402.01204, 2024.
- [Wang et al., 2024b] Zilong Wang, Hao Zhang, Chun-Liang Li, Julian Martin Eisenschlos, Vincent Perot, Zifeng Wang, Lesly Miculicich, Yasuhisa Fujii, Jingbo Shang, Chen-Yu Lee, et al. Chain-of-table: Evolving tables in the reasoning chain for table understanding. In *The Twelfth International Conference on Learning Representations*, 2024.

- [Wijmans and Baker, 1995] Johannes G Wijmans and Richard W Baker. The solution-diffusion model: a review. *Journal of membrane science*, 107(1-2):1–21, 1995.
- [Yan et al., 2023] Jiahuan Yan, Jintai Chen, Yixuan Wu, Danny Z Chen, and Jian Wu. T2g-former: organizing tabular features into relation graphs promotes heterogeneous feature interaction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 10720–10728, 2023.
- [Yang et al., 2022] Chengrun Yang, Gabriel Bender, Hanxiao Liu, Pieter-Jan Kindermans, Madeleine Udell, Yifeng Lu, Quoc V Le, and Da Huang. Tabnas: Rejection sampling for neural architecture search on tabular datasets. Advances in Neural Information Processing Systems, 35:11906–11917, 2022.
- [Yang et al., 2023] Yazheng Yang, Yuqi Wang, Guang Liu, Ledell Wu, and Qi Liu. Unitabe: Pretraining a unified tabular encoder for heterogeneous tabular data. arXiv preprint arXiv:2307.09249, 2023.
- [Yoon et al., 2020] Jinsung Yoon, Yao Zhang, James Jordon, and Mihaela van der Schaar. Vime: Extending the success of self-and semi-supervised learning to tabular domain. Advances in Neural Information Processing Systems, 33:11033–11043, 2020.
- [Zhang et al., 2023] Hengrui Zhang, Jiani Zhang, Balasubramaniam Srinivasan, Zhengyuan Shen, Xiao Qin, Christos Faloutsos, Huzefa Rangwala, and George Karypis. Mixedtype tabular data synthesis with score-based diffusion in latent space. arXiv preprint arXiv:2310.09656, 2023.
- [Zhang et al., 2024] Jinghan Zhang, Xiting Wang, Weijieying Ren, Lu Jiang, Dongjie Wang, and Kunpeng Liu. Ratt: Athought structure for coherent and correct Ilmreasoning. arXiv preprint arXiv:2406.02746, 2024.
- [Zhao et al., 2021] Zilong Zhao, Aditya Kunar, Robert Birke, and Lydia Y Chen. Ctab-gan: Effective table data synthesizing. In *Asian Conference on Machine Learning*, pages 97–112. PMLR, 2021.
- [Zhao et al., 2022a] Tianxiang Zhao, Enyan Dai, Kai Shu, and Suhang Wang. Towards fair classifiers without sensitive attributes: Exploring biases in related features. In Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining, pages 1433–1442, 2022.
- [Zhao *et al.*, 2022b] Tianxiang Zhao, Xiang Zhang, and Suhang Wang. Exploring edge disentanglement for node classification. In *Proceedings of the ACM Web Conference* 2022, pages 1028–1036, 2022.
- [Zhao et al., 2023] Tianxiang Zhao, Wenchao Yu, Suhang Wang, Lu Wang, Xiang Zhang, Yuncong Chen, Yanchi Liu, Wei Cheng, and Haifeng Chen. Skill disentanglement for imitation learning from suboptimal demonstrations. In Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pages 3513–3524, 2023.

- [Zhao et al., 2024] Tianxiang Zhao, Wenchao Yu, Suhang Wang, Lu Wang, Xiang Zhang, Yuncong Chen, Yanchi Liu, Wei Cheng, and Haifeng Chen. Interpretable imitation learning with dynamic causal relations. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 967–975, 2024.
- [Zhou et al., 2022] Kaixiong Zhou, Zirui Liu, Rui Chen, Li Li, S Choi, and Xia Hu. Table2graph: Transforming tabular data to unified weighted graph. In Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI, pages 2420–2426, 2022.
- [Zhu et al., 2020] Jiong Zhu, Yujun Yan, Lingxiao Zhao, Mark Heimann, Leman Akoglu, and Danai Koutra. Beyond homophily in graph neural networks: Current limitations and effective designs. Advances in neural information processing systems, 33:7793–7804, 2020.
- [Zhu et al., 2023] Bingzhao Zhu, Xingjian Shi, Nick Erickson, Mu Li, George Karypis, and Mahsa Shoaran. Xtab: Cross-table pretraining for tabular transformers. arXiv preprint arXiv:2305.06090, 2023.