Deep Neural Networks and Tabular Data: A Survey

Vadim Borisov, Tobias Leemann, Kathrin Seßler, Johannes Haug,
Martin Pawelczyk and Gjergji Kasneci

***Abstract*—Heterogeneous tabular data are the most commonly used form of data and are essential for numerous critical and computationally demanding applications. On homogeneous data sets, deep neural networks have repeatedly shown excellent performance and have therefore been widely adopted. However, their adaptation to tabular data for inference or data generation tasks remains highly challenging. To facilitate further progress in the field, this work provides an overview of state-of-the-art deep learning methods for tabular data. We categorize these methods into three groups: data transformations, specialized architectures, and regularization models. For each of these groups, our work offers a comprehensive overview of the main approaches. Moreover, we discuss deep learning approaches for generating tabular data, and we also provide an overview over strategies for explaining deep models on tabular data. Thus, our first contribution is to address the main research streams and existing methodologies in the mentioned areas, while highlighting relevant challenges and open research questions. Our second contribution is to provide an empirical comparison of traditional machine learning methods with eleven deep learning approaches across five popular real-world tabular data sets of different sizes and with different learning objectives. Our results, which we have made publicly available as competitive benchmarks, indicate that algorithms based on gradient-boosted tree ensembles still mostly outperform deep learning models on supervised learning tasks, suggesting that the research progress on competitive deep learning models for tabular data is stagnating. To the best of our knowledge, this is the first in-depth overview of deep learning approaches for tabular data; as such, this work can serve as a valuable starting point to guide researchers and practitioners interested in deep learning with tabular data.**

***Index Terms*—Deep neural networks, Tabular data, Heterogeneous data, Discrete data, Tabular data generation, Probabilistic modeling, Interpretability, Benchmark, Survey**

I. Introduction

Ever-increasing computational resources and the availability of large, labelled data sets have accelerated the success of deep neural networks [1], [2]. In particular, architectures based on convolutions, recurrent mechanisms [3], or transformers [4] have led to unprecedented performance in a multitude of domains. Although deep learning methods perform outstandingly well for classification or data generation tasks on homogeneous data (e.g., image, audio, and text data), tabular data still pose a challenge to deep learning models [5]–[8]. Tabular data – in contrast to image or language data – are heterogeneous, leading to dense numerical and sparse categorical features. Furthermore, the correlation among the features is weaker than the one introduced through spatial or semantic relationships in image or speech data. Hence, it is necessary to discover and exploit relations without relying on spatial information [9]. Therefore, Kadra et al. called tabular data sets the last *"unconquered castle"* for deep neural network models [10].

Heterogeneous data are the most commonly used form of data [7], and it is ubiquitous in many crucial applications, such as medical diagnosis based on patient history [11]–[13], predictive analytics for financial applications (e.g., risk analysis, estimation of creditworthiness, the recommendation of investment strategies, and portfolio management) [14], click-through rate (CTR) prediction [15], user recommendation systems [16], customer churn prediction [17], [18], cybersecurity [19], fraud detection [20], identity protection [21], psychology [22], delay estimations [23], anomaly detection [24], and so forth. In all these applications, a boost in predictive performance and robustness may have considerable benefits for both end users and companies that provide such solutions. Simultaneously, this requires handling many data-related pitfalls, such as noise, impreciseness, different attribute types and value ranges, or the missing value problem and privacy issues.

Meanwhile, deep neural networks offer multiple advantages over traditional machine learning methods. First, these methods are highly flexible [25], allow for efficient and iterative training, and are particularly valuable for AutoML [26]–[31]. Second, tabular data generation is possible using deep neural networks and can, for instance, help mitigate class imbalance problems [32]. Third, neural networks can be deployed for multimodal learning problems where tabular data can be one of many input modalities [28], [33]–[36], for tabular data distillation [37], [38], for federated learning [39], and in many more scenarios.

Successful deployments of data-driven applications require solving several tasks, among which we identified three *core challenges*: (1) *inference* (2) *data generation*, and (3) *interpretability*. The most crucial task is inference which is concerned with making predictions based on past observations. While a powerful predictive model is critical for all the applications mentioned in the previous paragraph, the interplay between tabular data and deep neural networks goes beyond simple inference tasks. Before a predictive model can even be trained, the training data usually needs to be preprocessed. This is where data generation plays a crucial role, as one of the standard deployment steps involves the imputation of missing values [40]–[42] and the rebalancing of the data set [43], [44] (i.e., equalizing sample sizes for different classes). Furthermore, it might be simply impossible to use

All authors are with the Data Science and Analytics Research (DSAR) group at the University of Tübingen, 72070 Tübingen, Germany. Gjergji Kasneci is also affiliated with Schufa Holding AG, 65201 Wiesbaden, Germany.
Corresponding authors:
vadim.borisov@uni-tuebingen.de
tobias.leemann@uni-tuebingen.de

the actual data due to privacy concerns, e.g., in financial or medical applications [45], [46]. Thus, to tackle the data preprocessing and privacy challenges, probabilistic tabular data *generation* is essential. Finally, with stricter data protection laws such as California Consumer Privacy Act (CCPA) [47] and the European General Data Protection Regulation (EU GDPR) [48], which both mandate a right to explanations for automated decision systems (e.g., in the form or recourse [49]), *interpretability* is becoming a key aspect for predictive models used for tabular data [50], [51]. During deployment, interpretability methods also serve as a valuable tool for model debugging and auditing [52].

Evidently, apart from the the core challenges of inference, generation, and interpretability, there are several other important subfields, such as working with data streams, distribution shifts, as well as privacy and fairness considerations that should not be neglected. Nevertheless, to navigate the vast body of literature, we focus on the identified core problems and thoroughly review the state of the art in this work. We will briefly discuss the remaining topics at the end of this survey.

Beyond reviewing current literature, we think that an exhaustive comparison between existing deep learning approaches for heterogeneous tabular data is necessary to put reported results into context. The variety of benchmarking data sets and the different setups often prevent comparison of results across papers. Additionally, important aspects of deep learning models, such as training and inference time, model size, and interpretability, are usually not discussed. We aim to bridge this gap by providing a comparison of the surveyed inference approaches with classical – yet very strong – baselines such as XGBoost [53]. We open-source our code, allowing researchers to reproduce and extend our findings.

In summary, the aims of this survey are to provide:

1) a thorough review of existing scientific literature on deep learning for tabular data;
2) a taxonomic categorization of the available approaches for classification and regression tasks on heterogeneous tabular data;
3) a presentation of the state of the art and promising paths towards tabular data generation;
4) an overview of existing explanation approaches for deep models for tabular data;
5) an extensive empirical comparison of traditional machine learning methods and deep learning models on multiple real-world heterogeneous tabular data sets;
6) a discussion on the main reasons for the limited success of deep learning on tabular data;
7) a list of open challenges related to deep learning for tabular data.

Accordingly, this survey is structured as follows: We discuss related works in Section II. To introduce the reader to the field, in Section III, we provide definitions of the key terms, a brief outline of the domain's history, and propose a unified taxonomy of current approaches to deep learning with tabular data. Section IV covers the main methods for modelling tabular data using deep neural networks. Section V presents an overview on tabular data generation using deep neural networks. An overview of explanation mechanisms for deep models for tabular data is presented in Section VI. In Section VII, we provide an extensive empirical comparison of machine and deep learning methods on real-world data, that also involves model size, runtime, and interpretability. In Section VIII, we summarize the state of the field and give future perspectives. Finally, we outline several open research questions before concluding in Section IX.

CONTENTS

II. Related Work

To the best of our knowledge, there is no study dedicated exclusively to the application of deep neural networks to tabular data, spanning the areas of supervised and unsupervised learning, data synthesis, and interpretability. Prior works cover some of these aspects, but none of them systematically discusses the existing approaches in the broadness of this survey.

However, there are some works that cover parts of the domain. There is a comprehensive analysis of common approaches for categorical data encoding as a preprocessing step for deep neural networks by Hancock & Khoshgoftaar [54]. The authors compared existing methods for categorical data encoding on various tabular data sets and different deep learning architectures. We also discuss the key categorical data encoding methods in Section IV-A1.

A recent survey by Sahakyan et al. [50] summarizes explanation techniques in the context of tabular data. Hence, we do not provide a detailed discussion of explainable machine learning for tabular data in this paper. However, for the sake of completeness, we present some of the most relevant works in Section VI and highlight open challenges in this area.

Gorishniy et al. [55] empirically evaluated a large number of state-of-the-art deep learning approaches for tabular data on a wide range of data sets. The authors demonstrated that a tuned deep neural network model with a ResNet-like architecture [56] shows comparable performance to some state-of-the-art deep learning approaches for tabular data.

Recently, Shwartz-Ziv & Armon [7] published a study on several different deep models for tabular data including TabNet [5], NODE [6], Net-DNF [57]. Additionally, they compared deep learning approaches to gradient boosting decision tree algorithms regarding accuracy, training effort, inference efficiency, and hyperparameter optimization time. They observed that deep models had the best results on their chosen data sets, however, not one single deep model could outperform all the others in general. The deep models were challenged by gradient boosting decision trees, leading the authors to conclude that efficient tabular data modelling using deep neural networks is still an open research problem. In the face of this evidence, we aim to integrate the necessary background for future research on the inference problem and on the intertwined challenges of generation and explainability into a single work.

III. Tabular Data and Deep Neural Networks

A. Definitions

In this section, we give definitions for central terms used in this work. We also provide pointers to the original works for more detailed explanations of the methods.

The key concept in this survey is a *(deep) neural network*. Unless stated otherwise we use this concept as a synonym for feed-forward networks, as described by [2], and name the concrete model whenever we deviate from this concept. A deep neural network defines a mapping $\hat{f}$,

$$y = f(x) \approx \hat{f}(x; W), \tag{1}$$

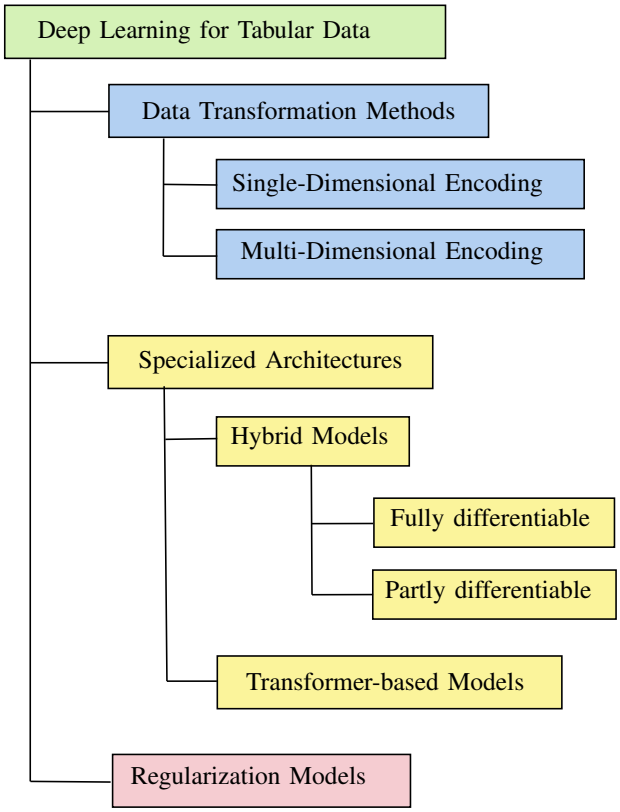


Fig. 1: Unified taxonomy of deep neural network models for heterogeneous tabular data.

that learns the value of the model parameters W (i.e., the "weights" of a neural network) that results in the best approximation of the true underlying and unknown function f. In this case, x is a multi-dimensional data sample (i.e., $x \in \mathbb{R}^n$) with corresponding target y (where typically, $y \in \mathbb{R}^k$ for k classes and $y \in \mathbb{R}$ for regression tasks) from a data set of tuples $\{(x_i, y_i)\}_{i \in \mathcal{I}}$. The network is called feed-forward if the input information flows in one direction to the output without any feedback connections.

Throughout this survey we focus on *heterogeneous data* which usually contains a variety of attribute types – These include both continuous and discrete attributes of different type (e.g., binary values, ordinal values, high-cardinality categorical values). This is fundamentally different from *homogeneous* data modalities, such as images, audio, or text data where only a single feature type is present.

Categorical variables are an attribute type of particular importance. According to Lane's definition [58], categorical variables are qualitative values. They "do not imply a numerical ordering", unlike quantitative values, which are "measured in terms of numbers". Usually, a categorical variable can take one out of a limited set of values. Examples of typical categorical variables include *gender*, *user_id*, *product_type* and *topic*.

Tabular data, sometimes also called *structured data* [59], is a subcategory of the heterogeneous data format that is usually presented in a table [60] with data points as rows and features as columns. In summary, for the scope of this work, we refer

Age	Education	Occupation	Sex	Income
39	Bachelors	Adm-clerical	Male	$\leq$50K
50	Bachelors	Exec-managerial	Male	$>$50K
38	HS-grad	Handlers-cleaners	Male	$\leq$50K
53	11th	Handlers-cleaners	Male	$\leq$50K
28	Bachelors	Prof-specialty	Female	$>$50K

TABLE I: An example of a heterogeneous tabular data set. Here we show five samples with selected variables from the Adult data set [61]. Section VII-A provides further details on this data set.

to a data set with a fixed number of features that are either continuous or categorical as tabular. Each data point can be understood as a row in the table, or – taking a probabilistic view – as a sample from the unknown joint distribution. An illustrative example of five rows of a heterogeneous, tabular data is provided in Table I.

B. A Brief History of Deep Learning on Tabular Data

Tabular data are one of the oldest forms of data to be statistically analysed. Before digital collection of text, images, and sound was possible, almost all data were tabular [62]–[64]. Therefore, it was the target of early machine learning research [65]. However, deep neural networks became popular in the digital age and were further developed with a focus on homogeneous data. In recent years, various supervised, self-supervised, and semi-supervised deep learning approaches have been proposed that explicitly address the issue of tabular data modelling again. Early works mostly focused on data transformation techniques for preprocessing [66]–[68], which are still important today [54].

A huge stimulus was the rise of e-commerce, which demanded novel solutions, especially in advertising [15], [69]. These tasks required fast and accurate estimation on heterogeneous data sets with many categorical variables, for which the traditional machine learning approaches are not well suited (e.g., categorical features that have high cardinality can lead to very sparse high-dimensional feature vectors and non-robust models). As a result, researchers and data scientists started looking for more flexible solutions, e.g., those based on deep neural networks, that can capture complex non-linear dependencies in the data.

In particular, the click-through rate prediction problem has received a lot of attention [15], [70], [71]. A large variety of approaches were proposed, most of them relying on specialized neural network architectures for heterogeneous tabular data.

A more recent line of research, sparked by Shavitt & Segal [72], evolved based on the idea that regularization may improve the performance of deep neural networks on tabular data [10]. This has led to an intensification of research on regularization approaches.

Due to the tremendous success of attention-based approaches such as transformers on textual [73] and visual data [74], [75], researchers have recently also started applying attention-based methods and self-supervised learning techniques to tabular data. After the introduction of transformer architectures to the field of tabular data [5], a lot of research effort has focused on transformer architectures that can be successfully applied to very large tabular data sets.

C. Challenges of Learning With Tabular Data

As we have mentioned in Section II, deep neural networks often perform less favourably compared to more traditional machine learning methods (e.g., tree-based methods) when dealing with tabular data. However, it is often unclear why deep learning cannot achieve the same level of predictive quality as in other domains such as image classification and natural language processing. In the following, we identify and discuss four possible reasons:

1) **Low-Quality Training Data**: Data quality is a common issue with real-world tabular data sets. They often include missing values [40], extreme data (outliers) [24], erroneous or inconsistent data [76], and have small overall size relative to the high-dimensional feature vectors generated from the data [77]. Also, due to the expensive nature of data collection, tabular data are frequently class-imbalanced. These challenges affect all machine learning algorithms; however, most of the modern decision tree-based algorithms can handle missing values or different/extreme variable ranges internally by looking for appropriate approximations and split values [53], [78], [79].
2) **Missing or Complex Irregular Spatial Dependencies:** There is often no spatial correlation between the variables in tabular data sets [80], or the dependencies between features are rather complex and irregular. When working with tabular data, the structure and relationships between its features have to be learned from scratch. Thus, the *inductive biases* used in popular models for homogeneous data, such as convolutional neural networks, are unsuitable for modelling this data type [57], [81], [82].
3) **Dependency on Preprocessing:** A key advantage of deep learning on homogeneous data is that it includes an implicit representation learning step [83], so only a minimal amount of preprocessing or explicit feature construction is required. However, for tabular data and deep neural networks the performance may strongly depend on the selected preprocessing strategy [84]. Handling the categorical features remains particularly challenging [54] and can easily lead to a very sparse feature matrix (e.g., by using a one-hot encoding scheme) or introduce a synthetic ordering of previously unordered values (e.g., by using an ordinal encoding scheme). Lastly, preprocessing methods for deep neural networks may lead to information loss, leading to a reduction in predictive performance [85].
4) **Importance of Single Features**: While typically changing the class of an image requires a coordinated change in many features, i.e., pixels, the smallest possible change of a categorical (or binary) feature can entirely flip a prediction on tabular data [72]. In contrast to deep neural networks, decision-tree algorithms can handle varying feature importance exceptionally well by selecting a

single feature and appropriate threshold (i.e., splitting) values and "ignoring" the rest of the data sample. Shavitt & Segal [72] have argued that individual weight regularization may mitigate this challenge and motivated more work in this direction [10].

With these four fundamental challenges in mind, we continue by organizing and discussing the strategies developed to address them. We start by developing a suitable taxonomy.

D. Unified Taxonomy

In this section, we introduce a taxonomy of approaches that allows for a unified view of the field. We divide the works from the deep learning with tabular data literature into three main categories: *data transformation methods*, *specialized architectures*, and *regularization models*. In Fig. 1, we provide an overview of our taxonomy of deep learning methods for tabular data.

Data transformation methods. The methods in the first group transform categorical and numerical data. This is usually done to enable deep neural network models to better extract the information signal. Methods from this group do not require new architectures or adaptations of the existing data processing pipeline. Nevertheless, the transformation step comes at the cost of an increased preprocessing time. This might be an issue for high-load systems [86], particularly in the presence of categorical variables with high cardinality and growing data set size. We can further subdivide this area into *Single-Dimensional Encodings* and *Multi-Dimensional Encodings*. The former encodings are employed to transform each feature independently while the latter encoding methods map an entire record to another representation.

Specialized architectures. The biggest share of works investigates specialized architectures and suggests that a different deep neural network architecture is required for tabular data. Two types of architectures are of particular importance: *hybrid models* fuse classical machine learning approaches (e.g., decision trees) with neural networks, while *transformer-based models* rely on attention mechanisms.

Regularization models. Lastly, the group of regularization models claims that one of the main reasons for the moderate performance of deep learning models on tabular data is their extreme non-linearity and model complexity. Therefore, strong regularization schemes are proposed as a solution. They are mainly implemented in the form of special-purpose loss functions.

We believe our taxonomy may help practitioners find the methods of choice that can be easily integrated into their existing tool chain. For instance, applying data transformations can result in performance improvements while maintaining the current model architecture. Conversely, using specialized architectures, the data preprocessing pipeline can be kept intact.

IV. Deep Neural Networks for Tabular Data

In this section, we discuss the use of deep neural networks on tabular data for classification and regression tasks according to the taxonomy presented in the previous section. We provide an overview of existing deep learning approaches in this area of research in Table II and examine the three methodological categories in detail: data transformation methods (see Subsection IV-A), architecture-based methods (see Subsection IV-B), and regularization-based models (see Subsection IV-C).

A. Data Transformation Methods

Most traditional approaches for deep neural networks on tabular data fall into this group. Interestingly, data preprocessing plays a relatively minor role in computer vision, even though the field is currently dominated by deep learning solutions [2]. There are many different possibilities to transform tabular data, and each may have a different impact on the learning results [54].

1) Single-Dimensional Encoding: One of the critical obstacles for deep learning with tabular data are categorical variables. Since neural networks only accept real number vectors as inputs, these values must be transformed before a model can use them. Therefore, the first class of methods attempts to encode categorical variables in a way suitable for deep learning models.

Approaches in this group [54] are divided into *deterministic* techniques, which can be used before training the model, and more complicated *automatic* techniques that are part of the model architecture. There are many ways for deterministic data encoding; hence we restrict ourselves to the most common ones without the claim of completeness.

The simplest data encoding technique might be ordinal or label encoding. Every category is just mapped to a discrete numeric value, e.g., $\{\texttt{Apple}, \texttt{Banana}\}$ are encoded as $\{0, 1\}$. One drawback of this method may be that it introduces an artificial order to previously unordered categories. Another straightforward method that does not induce any order is the one-hot encoding. One additional column for each unique category is added to the data. Only the column corresponding to the observed category is assigned the value one, with the other values being zero. In our example, `Apple` could be encoded as `(1,0)` and `Banana` as `(0,1)`. In the presence of a diverse set of categories in the data, this method can lead to high-dimensional sparse feature vectors and exacerbate the "curse of dimensionality" problem.

Binary encoding limits the number of new columns by transforming the qualitative data into a numerical representation (as the label encoding does) and using the binary format of the number. Again the digits, are split into different columns, but there are only $\log(c)$ new columns if c is the number of unique categorical values. If we extend our example to three fruits, e.g., $\{\texttt{Apple}, \texttt{Banana}, \texttt{Pear}\}$, we only need two columns to represent them: `(01)`, `(10)`, `(11)`.

One approach that needs no extra columns and does not include any artificial order is the so-called leave-one-out encoding. It is based on the target encoding technique proposed in the work by [101], where every category is replaced with the mean of the target variable of that category. The leave-one-out encoding excludes the current row when computing the mean of the target variable to avoid overfitting. This approach is also used in the CatBoost framework [79], a state-of-the-art machine learning library for heterogeneous tabular data based on the gradient boosting algorithm [102].

	Method	Interpretability	Key Characteristics
Encoding	SuperTML [87]		Transform tabular data into images for CNNs
	VIME [88]		Self-supervised learning and contextual embedding
	IGTD [80]		Transform tabular data into images for CNNs
	SCARF [89]		Self-supervised contrastive learning
Architectures, Hybrid	Wide&Deep [90]		Embedding layer for categorical features
	DeepFM [15]		Factorization machine for categorical data
	SDT [91]	✓	Distill neural network into interpretable decision tree
	xDeepFM [92]		Compressed interaction network
	TabNN [93]		DNNs based on feature groups distilled from GBDT
	DeepGBM [70]		Two DNNs, distill knowlegde from decision tree
	NODE [6]		Differentiable oblivious decision trees ensemble
	NON [94]		Network-on-network model
	DNN2LR [95]		Calculate cross feature wields with DNNs for LR
	Net-DNF [57]		Structure based on disjunctive normal form
	Boost-GNN [96]		GNN on top decision trees from the GBDT algorithm
	SDTR [97]		Hierarchical differentiable neural regression model
Architectures, Transformer	TabNet [5]	✓	Sequential attention structure
	TabTransformer [98]	✓	Transformer network for categorical data
	SAINT [9]	✓	Attention over both rows and columns
	ARM-Net [99]		Adaptive relational modelling with multi-headgated attention network
	Non-Param. Transformer [100]		Process the entire dataset at once, use attention between data points
Regul.	RLN [72]	✓	Hyperparameters regularization scheme
	Regularized DNNs [10]		A ”cocktail” of regularization techniques

TABLE II: Overview of deep learning approaches for tabular data. We organize them in categories ordered chronologically inside the groups. The “Interpretability” column indicates whether the approach offers some form interpretability for the model’s decisions. The key characteristics of every model are summarized in the last column.

A different strategy is hash-based encoding. Every category is transformed into a fixed-size value via a deterministic hash function. The output size is not directly dependent on the number of input categories but can be chosen manually.

2) Multi-Dimensional Encoding: A first automatic encoding strategy is the VIME approach [88]. The authors propose a self- and semi-supervised deep learning framework for tabular data that trains an encoder in a self-supervised fashion by using two pretext tasks. Those tasks that are independent from the concrete downstream task which the predictor has to solve. The first task of VIME is called mask vector estimation; its goal is to determine which values in a sample are corrupted. The second task, i.e., feature vector estimation, is to recover the original values of the sample. The encoder itself is a simple multilayer perceptron. This automatic encoding makes use of the fact that there is often much more unlabelled than labelled data. The encoder learns how to construct an informative homogeneous representation of the raw input data. In the semi-supervised step, a predictive model, which is also a deep neural network model, is trained using the labelled and unlabelled data transformed by the encoder. For the encoder, a novel data augmentation method is used, corrupting an unlabelled data point multiple times with different masks. On the predictions from all augmented samples from one original data point, a consistency loss $\mathcal{L}_u$ can be computed that rewards similar outputs. Combined with a supervised loss $\mathcal{L}_s$ from the labelled data, the predictive model minimizes the final loss $\mathcal{L} = \mathcal{L}_s + \beta \cdot \mathcal{L}_u$. To summarize, the VIME network trains an encoder, which is responsible to transform the categorical and numerical features into a new homogeneous and informative representation. This transformed feature vector is used as an input to the predictive model. For the encoder itself, the categorical data can be transformed by a simple one-hot-encoding and binary encoding.

Another stream of research aims at transforming the tabular input into a more homogeneous format. Since the revival of deep learning, convolutional neural networks have shown tremendous success in computer vision tasks. Therefore, the work by [87] proposed the SuperTML method, which is a data conversion technique to transform tabular data into an image data format (2-d matrices), i.e., black-and-white images.

The image generator for tabular data (IGTD) by [80] follows an idea similar to SuperTML. The IGTD framework converts tabular data into images to make use of classical convolutional architectures. As convolutional neural networks rely on spatial dependencies, the transformation into images is optimized by minimizing the difference between the feature distance

ranking of the tabular data and the pixel distance ranking of the generated image. Every feature corresponds to one pixel, which leads to compact images with similar features close at neighbouring pixels. Thus, IGDTs can be used in the absence of domain knowledge. The authors show relatively solid results for data with strong feature relationships but the method may fail if the features are independent or feature similarities can not characterize the relationships. In their experiments, the authors used only gene expression profiles and molecular descriptors of drugs as data. This kind of data may lead to a favourable inductive bias, so the general viability of the approach remains unclear.

B. Specialized Architectures

Specialized architectures form the largest group of approaches for deep tabular data learning. Hence, in this group, the focus is on the development and investigation of novel deep neural network architectures designed specifically for heterogeneous tabular data. Guided by the types of available models, we divide this group into two sub-groups: Hybrid models (presented in IV-B1) and transformer-based models (discussed in IV-B2).

1) Hybrid Models: Most approaches for deep neural networks on tabular data are hybrid models. They transform the data and fuse successful classical machine learning approaches, often decision trees, with neural networks. We distinguish between fully differentiable models, that can be differentiated with respect to all their parameters and partly differentiable models.

Fully differentiable Models. The fully differentiable models in this category offer a valuable property: They permit end-to-end deep learning for training and inference by means of gradient descent optimizers. Thus, they allow for highly efficient implementations in modern deep learning frameworks that exploit GPU or TPU acceleration throughout the code.

Popov et al. [6] propose an ensemble of differentiable oblivious decision trees [103] – also known as the NODE framework for deep learning on tabular data. Oblivious decision trees use the same splitting function for all nodes on the same level and can therefore be easily parallelized. NODE is inspired by the successful CatBoost [79] framework. To make the whole architecture fully differentiable and benefit from end-to-end optimization, NODE utilizes the entmax transformation [104] and soft splits. In the original experiments, the NODE framework outperforms XGBoost and other GBDT models on many data sets. As NODE is based on decision tree ensembles, there is no preprocessing or transformation of the categorical data necessary. Decision trees are known to handle discrete features well. In the official implementation, strings are converted to integers using the leave-one-out encoding scheme. The NODE framework is widely used and provides a sound implementation that can be readily deployed.

Frosst & Hinton [91] contribut another model relying on soft decision trees (SDT) to make neural networks more interpretable. They investigated training a deep neural network first, before using a mixture of its outputs and the ground truth labels to train the SDT model in a second step. This also allows for semi-supervised learning with unlabelled samples that are labelled by the deep neural network and used to train a more robust decision tree along with the labelled data. The authors showed that training a neural model first increases accuracy over SDTs that are directly learned from the data. However, their distilled trees still exhibit a performance gap to the neural networks that were fitted in the initial step. Nevertheless, the model itself shows a clear relationship among different classes in a hierarchical fashion. It groups different categorical values based on the common patterns, e.g., the digits 8 and 9 from the MNIST data set [105]. To summarize, the proposed method allows for high interpretability and efficient inference, at the cost of slightly reduced accuracy.

Follow-up work [97] extends this line of research to heterogeneous tabular data and regression tasks and presents the soft decision tree regressor (SDTR) framework. The SDTR is a neural network which imitates a binary decision tree. Therefore, all neurons, like nodes in a tree, get the same input from the data instead of the output from previous layers. In the case of deep networks, the SDTR could not beat other state-of-the-art models, but it has shown promising results in a low-memory setting, where single tree models and shallow architectures were compared.

Katzir et al. [57] follow a related idea. Their Net-DNF builds on the observation that every decision tree is merely a form of a Boolean formula, more precisely a disjunctive normal form. They use this inductive bias to design the architecture of a neural network, which is able to imitate the characteristics of the gradient boosting decision trees algorithm. The resulting Net-DNF was tested for classification tasks on data sets with no missing values, where it showed results that are comparable to those of XGBoost [53]. However, the authors did not mention how to handle high-cardinality categorical data, as the used data sets contained mostly numerical and few binary features.

Linear models (e.g., linear and logistic regression) provide global interpretability but are inferior to complex deep neural networks. Usually, handcrafted feature engineering is required to improve the accuracy of linear models. Liu et al. [95] use a deep neural network to combine the features in a possibly non-linear way; the resulting combination then serves as input to the linear model. This enhances the simple model while still providing interpretability.

The work by Cheng et al. [90] proposes a hybrid architecture that consists of linear and deep neural network models – Wide&Deep. A linear model that takes single features and a wide selection of hand-crafted logical expressions on features as an input is enhanced by a deep neural network to improve the generalization capabilities. Additionally, Wide&Deep learns an n-dimensional embedding vector for each categorical feature. All embeddings are concatenated resulting in a dense vector used as input to the neural network. The final prediction can be understood as a sum of both models. A similar work by Guo et al. [106] proposes an embedding using deep neural networks for categorical variables.

Another contribution to the realm of Wide&Deep models is DeepFM [15]. The authors demonstrate that it is possible to replace the hand-crafted feature transformations with learned Factorization Machines (FMs) [107], leading to an improvement

of the overall performance. The FM is an extension of a linear model designed to capture interactions between features within high-dimensional and sparse data efficiently. Similar to the original Wide&Deep model, DeepFM also relies on the same embedding vectors for its "wide" and "deep" parts. In contrast to the original Wide&Deep model, however, DeepFM alleviates the need for manual feature engineering.

Lastly, Network-on-Network (NON) [94] is a classification model for tabular data, which focuses on capturing the intra-feature information efficiently. It consists of three components: a field-wise network consisting of one unique deep neural network for every column to capture the column-specific information, an across-field-network, which chooses the optimal operations based on the data set, and an operation fusion network, connecting the chosen operations allowing for non-linearities. As the optimal operations for the specific data are selected, the performance is considerably better than that of other deep learning models. However, the authors did not include decision trees in their baselines, the current state-of-the-art models on tabular data. Also, training as many neural networks as columns and selecting the operations on the fly may lead to a long computation time.

Partly differentiable Models. This subgroup of hybrid models aims at combining non-differentiable approaches with deep neural networks. Models from this group usually utilize decision trees for the non-differentiable part.

The DeepGBM model [70] combines the flexibility of deep neural networks with the preprocessing capabilities of gradient boosting decision trees. DeepGBM consists of two neural networks – CatNN and GBDT2NN. While CatNN is specialized to handle sparse categorical features, GBDT2NN is specialized to deal with dense numerical features.

In the preprocessing step for the CatNN network, the categorical data are transformed via an ordinal encoding (to convert the potential strings into integers), and the numerical features are discretized, as this network is specialized for categorical data. The GBDT2NN network distills the knowledge about the underlying data set from a model based on gradient boosting decision trees by accessing the leaf indices of the decision trees. This embedding based on decision tree leaves was first proposed by [108] for the random forest algorithm. Later, the same knowledge distillation strategy has been adopted for gradient boosting decision trees [109].

Using the proposed combination of two deep neural networks, DeepGBM has a strong learning capacity for both categorical and numerical features. Distinctively, the authors implemented and tested DeepGBM's online prediction performance, which is significantly higher than that of gradient boosting decision trees. On the downside, the leaf indices can be seen as meta categorical features since these numbers cannot be directly compared. Also, it is not clear how other data-related issues, such as missing values, different scaling of numeric features, and noise influence the predictions produced by the models.

The TabNN architecture, introduced by [93], is based on two principles: explicitly leveraging expressive feature combinations and reducing model complexity. It distills the knowledge from gradient boosting decision trees to retrieve feature groups; it clusters them and then constructs the neural network based on those feature combinations. Also, structural knowledge from the trees is transferred to provide an effective initialization. However, the construction of the network already takes different extensive computation steps of which one is only a heuristic to avoid an NP-hard problem. Overall, considering the construction challenges and that an implementation of TabNN was not provided, the practical use of the network seems limited.

In similar spirit to DeepGBM and TabNN, the work from [96] proposes using gradient boosting decision trees for the data prepossessing step. The authors show that a decision tree structure has the form of a directed graph. Thus, the proposed framework exploits the topology information from the decision trees using graph neural networks [110]. The resulting architecture is coined Boosted Graph Neural Network (BGNN). In multiple experiments, BGNN demonstrates that the proposed architecture is superior to existing solid competitors in terms of predictive performance and training time.

2) Transformer-based Models: Transformer-based approaches form another subgroup of model-based deep neural methods for tabular data. Inspired by the recent surge of interest in transformer-based methods and their successes on text and visual data [75], [111], researchers and practitioners have proposed multiple approaches using deep attention mechanisms [4] for heterogeneous tabular data.

TabNet [5] is one of the first transformer-based models for tabular data. Like a decision tree, the TabNet architecture comprises multiple subnetworks that are processed in a sequential hierarchical manner. According to [5], each subnetwork corresponds to one *decision step*. To train TabNet, each decision step (subnetwork) receives the current data batch as input. TabNet aggregates the outputs of all decision steps to obtain the final prediction. At each decision step, TabNet first applies a sparse feature mask [112] to perform soft instance-wise feature selection. The authors claim that the feature selection can save valuable resources, as the network may focus on the most important features. The feature mask of a decision step is trained using attentive information from the previous decision step. To this end, a *feature transformer* module decides which features should be passed to the next decision step and which features should be used to obtain the output at the current decision step. Some layers of the feature transformers are shared across all decision steps. The obtained feature masks correspond to local feature weights and can also be combined into a global importance score. Accordingly, TabNet is one of the few deep neural networks that offers different levels of interpretability by design. Indeed, experiments show that each decision step of TabNet tends to focus on a particular subdomain of the learning problem (i.e., one particular subset of features). This behaviour is similar to convolutional neural networks. TabNet also provides a decoder module that is able to preprocess input data (e.g., replace missing values) in an unsupervised way. Accordingly, TabNet can be used in a two-stage self-supervised learning procedure, which improves the overall predictive quality. One of the popular Python [113] frameworks for tabular data provides an efficient implementation of TabNet [114]. Recently, TabNet has also been investigated in the context of fair machine learning [115], [116].

Attention-based architectures offer mechanisms for interpretability, which is an essential advantage over many hybrid models. Figure 2 shows attentions maps of the TabNet model and KernelSHAP explanation framework on the Adult data set [61].

Another supervised and semi-supervised approach is introduced by Huang et al. [98]. Their TabTransformer architecture uses self-attention-based transformers to map the categorical features to a contextual embedding. This embedding is more robust to missing or noisy data and enables interpretability. The embedded categorical features are then together with the numerical ones fed into a simple multilayer perceptron. If, in addition, there is an extra amount of unlabelled data, unsupervised pre-training can improve the results, using masked language modelling or replace token detection. Extensive experiments show that TabTransformer matches the performance of tree-based ensemble techniques, showing success also when dealing with missing or noisy data. The TabTransformer network puts a significant focus on the categorical features. It transforms the embedding of those features into a contextual embedding which is then used as input for the multilayer perceptron. This embedding is implemented by different multi-head attention-based transformers, which are optimized during training.

ARM-net [99] is an adaptive neural network for relation modelling tailored to tabular data. The key idea of the ARM-net framework is to model feature interactions with combined features (feature crosses) selectively and dynamically by first transforming the input features into exponential space and then determining the interaction order and interaction weights adaptively for each feature cross. Furthermore, the authors propose a novel sparse attention mechanism to generate the interaction weights given the input data dynamically. Thus, users can explicitly model feature crosses of arbitrary orders with noisy features filtered selectively.

SAINT (Self-Attention and Intersample Attention Transformer) [9] is a hybrid attention approach, combining self-attention [4] with inter-sample attention over multiple rows. When handling missing or noisy data, this mechanism allows the model to borrow the corresponding information from similar samples, which improves the model's robustness. The technique is reminiscent of nearest-neighbour classification. In addition, all features are embedded into a combined dense latent vector, enhancing existing correlations between values from one data point. To exploit the presence of unlabelled data, a self-supervised contrastive pre-training can further improve the results, minimizing the distance between two views of the same sample and maximizing the distance between different ones. Like the VIME framework (Section IV-A1), SAINT uses CutMix [117] to augment samples in the input space and uses mixup [118] in the embedding space.

Finally, even some new learning paradigms are being proposed. For instance, the Non-Parametric Transformer (NPT) [100] does not construct a mapping from individual inputs to outputs but uses the entire data set at once. By using attention between data points, relations between arbitrary samples can be modelled and leveraged for classifying test samples.

C. Regularization Models

The third group of approaches argues that extreme flexibility of deep learning models for tabular data is one of the main learning obstacles and strong regularization of learned parameters may improve the overall performance.

One of the first methods in this category was the Regularization Learning Network (RLN) proposed by Shavitt & Segal [72], which uses a learned regularization scheme. The main idea is to apply trainable regularization coefficients to each single weight in a neural network, thereby lowering the sensitivity. To efficiently determine the corresponding coefficients, the authors propose a novel loss function termed "Counterfactual Loss". The regularization coefficients lead to a very sparse network, which also provides the importance of the remaining input features.

In their experiments, RLNs outperform deep neural networks and obtain results comparable to those of the gradient boosting decision trees algorithm, but the evaluation relies on a data set with mainly numerical data to compare the models. The RLN paper does not address the issues of categorical data. For the experiments and the example implementation data sets with exclusively numerical data (except for the gender attribute) were used. A similar idea is proposed in [119], where regularization coefficients are learned only in the first layer with a goal to extract feature importance.

Kadra et al. [10] state that simple multilayer perceptrons can outperform state-of-the-art algorithms on tabular data if deep learning networks are properly regularized. The authors propose a "cocktail" of regularization with thirteen different techniques that are applied jointly. From those, the optimal subset and their subsidiary hyperparameters are selected. They demonstrate in extensive experiments that the "cocktails" regularization can not only improve the performance of multilayer perceptrons, but these simple models also outperform tree-based architectures. On the downside, the extensive per-data set regularization and hyperparameter optimization take much more computation time than the gradient boosting decision trees algorithm.

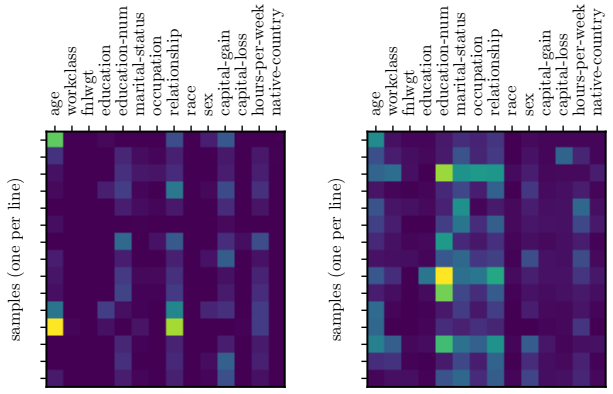
There are several other works [120]–[122] showing that strong regularization of deep neural networks can be beneficial for tabular data.

V. Tabular Data Generation

For many applications, the generation of realistic tabular data is fundamental. Two of the main purposes are data augmentation [124] and data imputation (i.e., the filling of missing values) [41], [42] and rebalancing [43], [44], [125], [126]. Another highly relevant topic is privacy-aware machine learning [45], [46], [127] where generated data can potentially be leveraged to overcome privacy concerns.

A. Methods

While the generation of images and text is highly explored [128]–[130], generating synthetic tabular data is still a challenge. The mixed structure of discrete and continuous features along with their different value distributions still poses a significant challenge.



(a) TabNet attributions (b) KernelSHAP attributions

Fig. 2: Interpretable learning with the TabNet [5] architecture. We compare the attributions provided by the model for a sample from the UCI Adult data set with those provided by the game theoretic KernelSHAP framework [123].

Classical approaches for the data generation task include Copulas [131], [132] and Bayesian Networks [133]. Among Bayesian Networks those based on the Chow-Liu approximation [134] are especially popular.

In the deep-learning era, Generative Adversarial Networks (GANs) [135] have proven highly successful for the generation of images [128], [136]. GANs were recently introduced as an original way to train a generative deep neural network model. They consist of two separate models: a generator G that generates samples from the data distribution, and a discriminator D that estimates the probability that a sample came from the ground truth distribution. Both G and D are usually chosen to be non-linear functions such as a multilayer perceptrons. To learn a generator distribution p_g over data $\boldsymbol{x}$, the generator $G(\boldsymbol{z}; \theta_g)$ maps samples from a noise distribution $p_z(\boldsymbol{z})$ (e.g., the Gaussian distribution) to the input data space. The discriminator $D(\boldsymbol{x}; \theta_d)$ outputs the probability that a data point $\boldsymbol{x}$ comes from the training data's distribution p_{data} rather than from the generator's output distribution p_g. During joint training of G and D, G will start generating successively more realistic samples to fool the discriminator D. For more details on GANs, we refer the interested reader to the original paper [135].

Although it was found that GANs lag behind at the generation of discrete outputs such as natural language [130], they are still frequently chosen to generate tabular data. Vanilla GANs or derivates such as the Wasserstein GAN (WGAN) [137], WGAN with gradient penalty (WGAN-GP) [138], Cramér GAN [139], or the Boundary seeking GAN [140], which is designed to model discrete data, are commonly used in the literature to generate tabular data. Moreover, VeeGAN [141] is frequently used for tabular data. Apart from GANs, autoencoder-based architectures – in particular those relying on Variational Autoencoders (VAEs) [142] – have been proposed [143], [144].

In Table III, we provide an overview of tabular generation approaches, that use deep learning techniques. Note that due to the enormous number of approaches, we list the most influential works that address the problem of data generation with a particular focus on tabular data. We exclude works that are targeted towards highly domain-specific tasks.

In the following section, we will briefly discuss the most relevant approaches that helped shape the domain. For example, MedGAN by [46] was one of the first works and provides a deep learning model to generate patient records. As all the features in their work are discrete, this model cannot be easily transferred to arbitrary tabular data sets. The table-GAN approach by [145] adapts the Deep Convolutional GAN for tabular data. Specifically, the features from one record are converted into a matrix, so that they can be processed by convolutional filters of a convolutional neural network. However, it remains unclear to which extent the inductive bias used for images are suitable for tabular data.

The approach by Xu et al. [144] focuses on the correlation between the features of one data point. The authors first propose the *mode-specific normalization* technique for data preprocessing that allows to transform non-Gaussian distributions in the continuous columns. They express numeric values in terms of a mixture component number and the deviation from that component's center. This allows to represent multi-modal and skewed distributions. Their generative solution, coined CTGAN, uses the conditional GAN architecture to enforce learning proper conditional distributions for each column. To obtain categorical values and to allow for backpropagation in the presence of categorical values, the *gumbel-softmax trick* [146] is utilized. The authors also propose a model based on Variational Autoencoders, named TVAE (Tabular Variational Autoencoder) which outperforms their suggested GAN approach. Both approaches can be considered state-of-the-art.

While GANs and VAEs are prevalent, other recently proposed architectures include machine-learned Causal Models [147] and Invertible Flows [45]. When privacy is the main factor of concern, models such as PATE-GAN [148] provide generative models with certain differential privacy guarantees. Although very relevant for practical applications, such privacy guarantees and related federated learning approaches with tabular data are outside the scope of this review.

Fan et al. [127] compare a variety of different GAN architectures for tabular data synthesis and recommend using a simple, fully connected architecture with a vanilla GAN loss with minor changes to prevent mode-collapse. They also use the normalization proposed by [144]. In their experiments, the Wasserstein GAN loss or the use of convolutional architectures on tabular data does boost the generative performance.

B. Assessing Generative Quality

To assess the quality of the generated data, several performance measures are used. The most common approach is to define a proxy classification task and train one model for it on the real training set and another on the artificially generated data set. With a highly capable generator, the predictive performance of the artificial-data model on the real-data test set should be almost on par with its real-data

Method	Based upon	Application
medGAN [46]	Autoencoder+GAN	Medical Records
TableGAN [145] DCGAN	General	
Mottini et al. [149]	Cramér GAN	Passenger Records
Camino et al. [150]	medGAN, ARAE	General
medBGAN, medWGAN [151]	WGAN-GP, Boundary seeking GAN	Medical Records
ITS-GAN [124]	GAN with AE for constraints	General
CTGAN, TVAE [144]	Wasserstein GAN, VAE	General
actGAN [126]	WGAN-GP	Health Data
VAEM [143]	VAE (Hierarchical)	General
OVAE [152]	Oblivious VAE	General
TAEI [44]	AE+SMOTE (in multiple setups)	General
Causal-TGAN [153]	Causal-Model, WGAN-GP	General
Copula-Flow [45]	Invertible Flows	General

TABLE III: Generation of tabular data using deep neural network models (in chronological order).

counterpart. This measure is often referred to as *machine learning efficacy* and used in [46], [144], [149]. In non-obvious classification tasks, an arbitrary feature can be used as a label and predicted [46], [150], [151]. Another approach is to visually inspect the modelled distributions per-feature, e.g., the cumulative distribution functions [124] or compare the expected values in scatter plots [46], [150]. A more quantitative approach is the use of statistical tests, such as the Kolmogorov-Smirnov test [154], to assess the distributional difference [151]. On synthetic data sets, the output distribution can be compared to the ground truth, e.g., in terms of log-likelihood [144], [147]. Because overfitted models can also obtain good scores, [144] propose evaluating the likelihood of a test set under an estimate of the GAN's output distribution. Especially in a privacy-preserving context, the distribution of the *Distance to Closest Record* (DCR) can be calculated and compared to the respective distances on the test set [145]. This measure is important to assess the extent of sample memorization. Overall, we conclude that a single measure is not sufficient to assess the generative quality. For instance, a generative model that memorizes the original samples will score well in the machine learning efficiency metric but fail the DCR check. Therefore, we highly recommend using several evaluation measures that focus on individual aspects of data quality.

VI. Explanation Mechanisms for Deep Learning with Tabular Data

Explainable machine learning is concerned with the problem of providing explanations for complex machine learning models. With stricter regulations for automated decision making [48] and the adoption of machine learning models in high-stakes domains such as finance and healthcare [52], interpretability is becoming a key concern. Towards this goal, various streams of research follow different explainability paradigms. Among these, feature attribution methods and counterfactual explanations are two of the popular forms [155]–[157]. Because these techniques are gaining importance for researchers and practitioners alike, we dedicate the following section to reviewing these methods.

A. Feature Highlighting Explanations

Local input attribution techniques seek to explain the behaviour of machine learning models instance by instance. Those methods aim to highlight the influence the inputs have on the prediction by assigning importance scores to the input features. Some popular approaches for model explanations aim at constructing classification models that are explainable by design [158]–[160]. This is often achieved by enforcing the deep neural network model to be locally linear. Moreover, if the model's parameters are known and can be accessed, then the explanation technique can use these parameters to generate the model explanation. For such settings, relevance-propagation-based methods, e.g., [161], [162], and gradient-based approaches, e.g., [163]–[165], have been suggested. In cases where the parameters of the neural network cannot be accessed, model-agnostic approaches can prove useful. This group of approaches seeks to explain a model's behavior locally by applying surrogate models [123], [166]–[169], which are interpretable by design and are used to explain individual predictions of black-box machine learning models. In order to test the performance of these black-box explanations techniques, Liu et al. [170] suggest a python-based benchmarking library.

B. Counterfactual Explanations

From the perspective of algorithmic recourse, the main purpose of counterfactual explanations is to suggest constructive interventions to the input of a deep neural network so that the output changes to the advantage of an end user. In simple terms, a minimal change to the feature vector that will flip the classification outcome is computed and provided as an explanation. By emphasizing both the feature importance and

the recommendation aspect, counterfactual explanation methods can be further divided into three different groups: works that assume that all features can be independently manipulated [171] and works that focus on manifold constraints to capture feature dependencies.

In the class of independence-based methods, where the input features of the predictive model are assumed to be independent, some approaches use combinatorial solvers to generate recourse in the presence of feasibility constraints [172]–[175]. Another line of research deploys gradient-based optimization to find low-cost counterfactual explanations in the presence of feasibility and diversity constraints [176]–[178]. The main problem with these approaches is that they abstract from input correlations.

To alleviate this problem and to suggest realistic looking counterfactuals, researchers have suggested building recourse suggestions on generative models [179]–[184]. The main idea is to change the geometry of the intervention space to a lower dimensional latent space, which encodes different factors of variation while capturing input dependencies. To this end, these methods primarily use (tabular data) variational autoencoders [142], [185]. In particular, Mahajan et al. [182] demonstrate how to encode various feasibility constraints into such models. However, an extensive comparison across this class of methods is still missing since it is difficult to measure how realistic the generated data are in the context of algorithmic recourse.

More recently, a few works have suggested to develop counterfactual explanations that are robust to model shifts and noise in the recourse implementations [186]–[188]. A comprehensive treatment on how to extend these lines of work to arbitrary high cardinality categorical variables is still an open problem in the field.

For a more fine-grained overview over the literature on counterfactual explanations we refer the interested reader to the most recent surveys [189], [190]. Finally, Pawelczyk et al. [157] have implemented an open-source python library which provides support for many of the aforementioned counterfactual explanation models.

VII. Experiments

Although several experimental studies have been published in recent years [7], [10], an exhaustive comparison between existing deep learning approaches for heterogeneous tabular data is still missing in the literature. For example, important aspects of deep learning models such as training and inference time, model size, and interpretability, are not discussed.

To fill this gap, we present an extensive empirical comparison of machine and deep learning methods on real-world data sets with varying characteristics in this section. We discuss the data set choice (VII-A), the results (VII-B), and present a comparison of the training and inference time for all the machine learning models considered in this survey (VII-C). We also discuss the size of deep learning models. Lastly, to the best of our knowledge, we present the first comparison of explainable deep learning methods for tabular data (VII-D). We release the full source code of our experiments for maximum transparency[1].

[1]Open benchmarking on tabular data for machine learning models: https://github.com/kathrinse/TabSurvey.

A. Data Sets

In computer vision, there are many established data sets for the evaluation of new deep learning architectures such as MNIST [105], CIFAR [191], and ImageNet [192]. On the contrary, there are no established standard heterogeneous data sets. Carefully checking the works listed in Section IV, we identified over 100 different data sets with different characteristics in their respective experimental evaluation sections. We note that the small overlap between the mentioned works makes it hard to compare the results across these works in general. Therefore, in this work, we deliberately select data sets covering the entire range of characteristics, such as data domain (e.g., finance, e-commerce, geography, physics), different types of target variables (classification, regression), varying number of categorical variables and continuous variables, and differing sample sizes (small to large). Furthermore, most of the selected data sets were previously featured in multiple studies.

The first data set of our study is the Home Equity Line of Credit (HELOC) data set provided by FICO [193]. This data set consists of anonymized information from real homeowners who applied for home equity lines of credit. A HELOC is a line of credit typically offered by a bank as a percentage of home equity. The task consists of using the information about the applicant in their credit report to predict whether they will repay their HELOC account within a two-year period.

We further use the Adult Income data set [61], which is among the most popular tabular data sets used in the surveyed work (5 usages). It includes basic information about individuals such as: age, gender, education, etc. The target variable is binary; it represents high and low income.

The largest tabular data set in our study is HIGGS, which stems from particle physics. The task is to distinguish between signals with Higgs bosons (HIGGS) and a background process [194]. Monte Carlo simulations [195] were used to produce the data. In the first 21 columns (columns 2-22), the particle detectors in the accelerator measure kinematic properties. In the last seven columns, these properties are analyzed. In total, HIGGS includes eleven million rows. In contrast to other data sets of our study, the HIGGS data set contains only numerical or continuous variables. Since DeepFM, DeepGBM, and TabTransformer models require at least one categorical attribute, we binarize the twenty-first variable into a categorical variable with three groups.

The Covertype data set [61] is multi-classification data set which holds cartographic information about land cells (e.g., elevation, slope). The goal is to predict which one out of seven forest cover types is present in the cell.

Finally, we utilize the California Housing data set [196], which contains information about a number of properties. The prediction task (regression) is to estimate price of the corresponding home.

The fundamental characteristics of the selected data sets are summarized in Table VI.

B. Open Performance Benchmark on Tabular Data

1) Hyperparameter Selection: In order to do a fair evaluation, we use the Optuna library [203] with 100 iterations

	HELOC	Adult Income	HIGGS	Covertype	California Housing
Samples	9.871	32.561	11 M.	581.012	20.640
Num. features	21	6	27	52	8
Cat. features	2	8	1	2	0
Task	Binary	Binary	Binary	Multi-Class	Regression
Classes	2	2	2	7	-

TABLE IV: Main properties of the real-world heterogeneous tabular data sets used in this survey. We also indicate the data set task, where "Binary" stands for binary classification, and "Multi-class" represents multi-class classification.

for each model to tune hyperparameters. Each hyperparameter configuration was cross-validated with five folds. The hyperparameter ranges used are publicly available online along with our code. We laid out the search space based on the information given in the corresponding papers and recommendations from the framework's authors.

2) Data Preprocessing: We prepossessed the data in the same way for every machine learning model by applying zero-mean, unit-variance normalization to the numerical features and an ordinal encoding to the categorical ones. The missing values were substituted with zeros for the linear regression and models based on pure neural networks since these methods cannot accept them otherwise. We apply the ordinal encoding to categorical values for all models. According to the work [54], the chosen encoding strategy shows comparable performance to more advanced methods. We explicitly specify which features are categorical for LightGBM, DeepFM, DeepGBM, TabNet, TabTransformer, and SAINT, since these approaches provide special functionality dedicated to categorical values, e.g., learning an embedding of the categories.

3) Reproducibility and Extensibility: For maximum reproducibility, we run all experiments in a docker container [204]. We underline again that our full code is publicly released so that the experiments can be replicated. The mentioned datasets are also publicly available and can be used as a benchmark for novel methods. We would highly welcome contributed implementations of additional methods from the data science community.

4) Results: The results of our experiments are shown in Table V. They draw a different picture than many recent research papers may suggest: For all but the very large HIGGS data set, the best scores are still obtained by boosted decision tree ensembles. XGBoost and CatBoost outperform all deep learning-based approaches on the small and medium data sets, the regression data set, and the multi-class data set. For the large-scale HIGGS, SAINT outperforms the classical machine learning approaches. This suggests that for very large tabular data sets with predominantly continuous features, modern neural network architectures may have an advantage over classical approaches after all. In general, however, our results are consistent with the inferior performance of deep learning techniques in comparison to approaches based on decision tree ensembles (such as gradient boosting decision trees) on tabular data that was observed in various Kaggle competitions [205].

Considering only deep learning approaches, we observe that SAINT provided competitive results across data sets. However, for the other models, the performance was highly dependent on the chosen data set. DeepFM performed best (among the deep learning models) on the Adult dataset and second-best on the California Housing data set, but returned only weak results on HELOC.

C. Run Time Comparison

We also analyse the training and inference time of the models in comparison to their performance. We plot the time-performance characteristic for the models in Fig. 3 and Fig. 4 for the Adult and the HIGGS data set respectively. While the training time of gradient-boosting-based models is lower than that of most deep neural network-based methods, their inference time on the HIGGS data set with 11 million samples is significantly higher: for XGBoost, the inference time amounts to 5995 seconds whereas inference times for MLP and SAINT are 10.18 and 282 seconds respectively. All gradient-boosting and deep learning models were trained on the same GPU.

D. Interpretability Assessment

As opposed to the pure on-task performance, interpretability of the models is becoming an increasingly important characteristic. Therefore, we end this section with a distinct assessment of the interpretability properties claimed by some methods. The model size (number of parameters) can provide a first intuition of the interpretability of the models. Therefore, we provide a size comparison of deep learning models in Fig. 5.

Admittedly, explanations can be provided in very different forms, which may each have their own use-cases. Hence, we can only compare explanations that have a common form. In this work, we chose feature attributions as the explanation format because they are the prevalent form of post-hoc explainability for the models considered in this work. Remarkably, the models that build on the transformer architecture (Section IV-B2) often claim some extent of interpretability through the attention maps [9]. To verify this hypothesis and assess the attribution provided by some of the frameworks in practice, we run an ablation test with the features that were attributed the highest importance over all samples. Furthermore, due to the lack of ground truth attribution values, we compare individual attributions to the well-known KernelSHAP values [123].

Evaluation of the quality of feature attribution is known to be a non-trivial problem [206]. We measure the fidelity [207] of the attributions by successively removing the features that have the highest mean importance assigned (Most Relevant First, MoRF [207]). We then retrain the model on the reduced feature set. A sharp drop in predictive accuracy indicates that the discriminative features were successfully identified and removed. We do the same for the inverse order, Least Relevant First (LeRF), which removes the features deemed unimportant. In this case, the accuracy should stay high as long as possible. For the attention maps of TabTransformer and SAINT, we either use the sum over the entire columns of the intra-feature attention maps as an importance estimate or only take the diagonal (feature self-attentions) as attributions.

The obtained curves are visualized in Fig. 6. For the MoRF order, TabNet and TabTransformer with the diagonal of the

	HELOC		Adult		HIGGS		Covertype		Cal. Housing
	Acc ↑	AUC ↑	Acc ↑	AUC ↑	Acc ↑	AUC ↑	Acc ↑	AUC ↑	MSE ↓
Linear Model	73.0±0.0	80.1±0.1	82.5±0.2	85.4±0.2	64.1±0.0	68.4±0.0	72.4±0.0	92.8±0.0	0.528±0.008
KNN [65]	72.2±0.0	79.0±0.1	83.2±0.2	87.5±0.2	62.3±0.1	67.1±0.0	70.2±0.1	90.1±0.2	0.421±0.009
Decision Tree [197]	80.3±0.0	89.3±0.1	85.3±0.2	89.8±0.1	71.3±0.0	78.7±0.0	79.1±0.0	95.0±0.0	0.404±0.007
Random Forest [198]	82.1±0.2	90.0±0.2	86.1±0.2	91.7±0.2	71.9±0.0	79.7±0.0	78.1±0.1	96.1±0.0	0.272±0.006
XGBoost [53]	83.5±0.2	92.2±0.0	87.3±0.2	92.8±0.1	77.6±0.0	85.9±0.0	**97.3±0.0**	**99.9±0.0**	0.206±0.005
LightGBM [78]	83.5±0.1	92.3±0.0	**87.4±0.2**	**92.9±0.1**	77.1±0.0	85.5±0.0	93.5±0.0	99.7±0.0	**0.195±0.005**
CatBoost [79]	**83.6±0.3**	**92.4±0.1**	87.2±0.2	92.8±0.1	77.5±0.0	85.8±0.0	96.4±0.0	99.8±0.0	0.196±0.004
Model Trees [199]	82.6±0.2	91.5±0.0	85.0±0.2	90.4±0.1	69.8±0.0	76.7±0.0	-	-	0.385±0.019
MLP [200]	73.2±0.3	80.3±0.1	84.8±0.1	90.3±0.2	77.1±0.0	85.6±0.0	91.0±0.4	76.1±3.0	0.263±0.008
DeepFM [15]	73.6±0.2	80.4±0.1	86.1±0.2	91.7±0.1	76.9±0.0	83.4±0.0	-	-	0.260±0.006
DeepGBM [70]	78.0±0.4	84.1±0.1	84.6±0.3	90.8±0.1	74.5±0.0	83.0±0.0	-	-	0.856±0.065
RLN [72]	73.2±0.4	80.1±0.4	81.0±1.6	75.9±8.2	71.8±0.2	79.4±0.2	77.2±1.5	92.0±0.9	0.348±0.013
TabNet [5]	81.0±0.1	90.0±0.1	85.4±0.2	91.1±0.1	76.5±1.3	84.9±1.4	93.1±0.2	99.4±0.0	0.346±0.007
VIME [88]	72.7±0.0	79.2±0.0	84.8±0.2	90.5±0.2	76.9±0.2	85.5±0.1	90.9±0.1	82.9±0.7	0.275±0.007
TabTransformer [98]	73.3±0.1	80.1±0.2	85.2±0.2	90.6±0.2	73.8±0.0	81.9±0.0	76.5±0.3	72.9±2.3	0.451±0.014
NODE [6]	79.8±0.2	87.5±0.2	85.6±0.3	91.1±0.2	76.9±0.1	85.4±0.1	89.9±0.1	98.7±0.0	0.276±0.005
Net-DNF [57]	82.6±0.4	91.5±0.2	85.7±0.2	91.3±0.1	76.6±0.1	85.1±0.1	94.2±0.1	99.1±0.0	-
STG [201]	73.1±0.1	80.0±0.1	85.4±0.1	90.9±0.1	73.9±0.1	81.9±0.1	81.8±0.3	96.2±0.0	0.285±0.006
NAM [202]	73.3±0.1	80.7±0.3	83.4±0.1	86.6±0.1	53.9±0.6	55.0±1.2	-	-	0.725±0.022
SAINT [9]	82.1±0.3	90.7±0.2	86.1±0.3	91.6±0.2	**79.8±0.0**	**88.3±0.0**	96.3±0.1	99.8±0.0	0.226±0.004

TABLE V: Open performance benchmark results based on (stratified) 5-fold cross-validation. We use the same fold splitting strategy for every data set. The top results for each dataset are in **bold**, we also underline the second-best results. The mean and standard deviation values are reported for each baseline model. Missing results indicate that the corresponding model could not be applied to the task type (regression or multi-class classification).

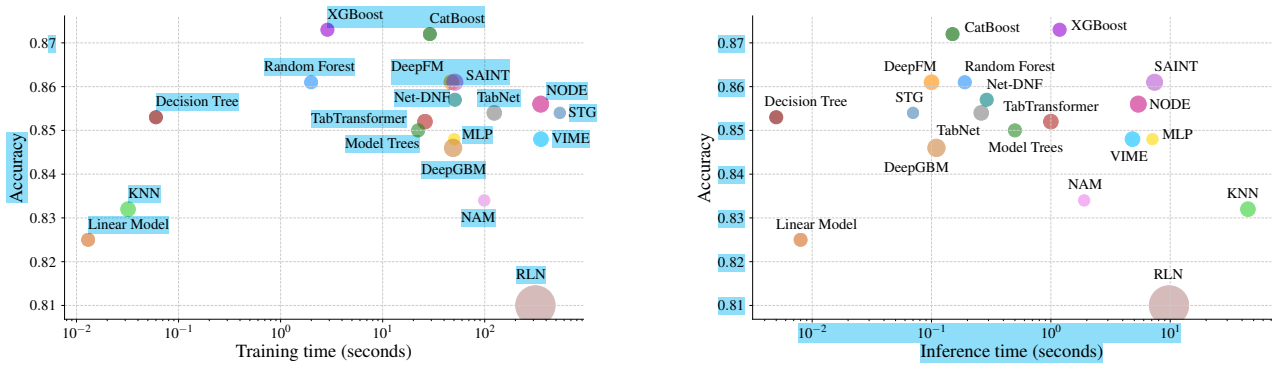


Fig. 3: Train (left) and inference (right) time benchmarks for selected methods on the Adult data set with 32.561 samples. The circle size reflects the accuracy standard deviation.

attention head as attributions seem to perform best. For LeRF, TabNet is the only significantly better method than the others. For TabTransformer, taking the diagonal of the attention matrix seems to increase the performance, whereas for SAINT, there is almost no difference. We additionally compare the attribution values obtained to values from the KernelSHAP attribution method. Unfortunately, there are no ground truth attributions to compare with. However, the SHAP framework has a solid grounding in game theory and is widely deployed [50]. We only compare the absolute values of the attributions, as the attention maps are constrained to be positive. As a measure of agreement, we compute the Spearman Rank Correlation between the attributions by the SHAP framework and the tabular data models. The correlation we observe is surprisingly low across all models, and sometimes it is even negative, which means that a higher SHAP attribution will probably result in a lower attribution by the model.

In these two simple benchmarks, the transformer models were not able to produce convincing feature attributions out-of-the-box. We come to the conclusion that more profound benchmarks of the claimed interpretability characteristics and their usefulness in practice are necessary.

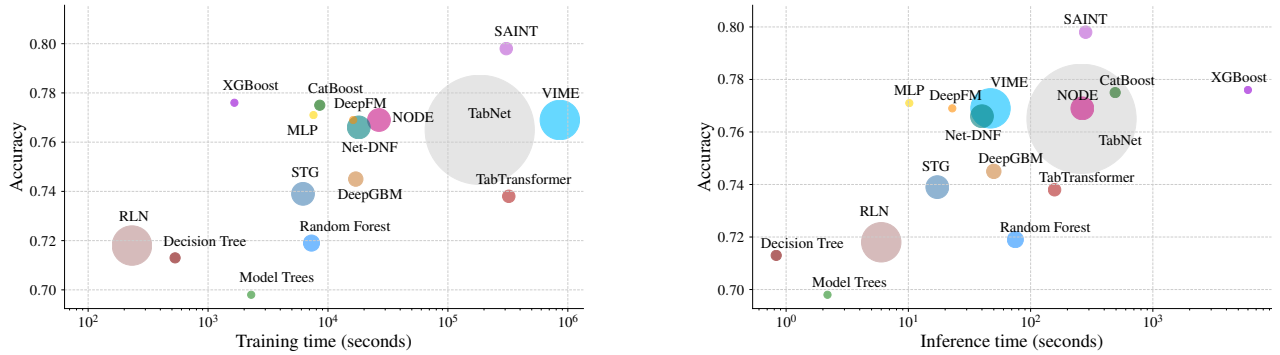


Fig. 4: Train (left) and inference (right) time benchmarks for selected methods on the HIGGS data set with 11 million samples. The circle size reflects the accuracy standard deviation.

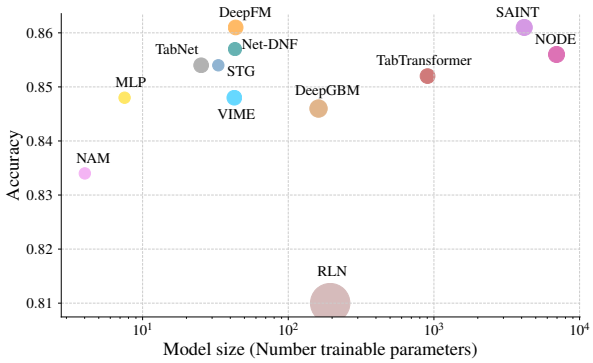
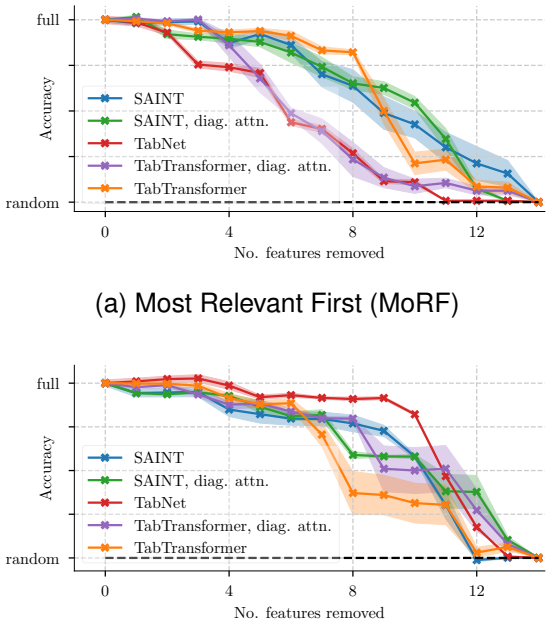


Fig. 5: A size comparison of deep learning models on the Adult data set. The circle size reflects standard deviation.

Model, attention used	Spearman Corr.
TabTransformer, columnw. attention	-0.01 ± 0.008
TabTransformer, diag. attention	0.00 ± 0.010
TabNet	**0.07** ± 0.009
SAINT, columnw. attention	-0.04 ± 0.007
SAINT, diag. attention	0.01 ± 0.007

TABLE VI: Spearman rank correlation of the provided attribution with KernelSHAP values as ground truth. Results were computed on 750 random samples from the Adult data set.



(a) Most Relevant First (MoRF)

(b) Least Relevant First (LeRF)

Fig. 6: Resulting curves of the global attribution benchmark for feature attributions (15 runs on Adult). Standard errors are indicated by the shaded area. For the MoRF order, an early drop in accuracy is desirable, while for LeRF, the accuracy should stay as high as possible.

VIII. Discussion and Future Prospects

In this section, we summarize our findings and discuss current and future trends in deep learning approaches for tabular data (Section VIII-A). Moreover, we identify several open research questions that could be tackled to advance the field of tabular deep neural networks (Section VIII-B).

A. Summary and Trends

Decision Tree Ensembles are still State-of-the-Art. In a fair comparison on multiple data sets, we demonstrated that models based on tree-ensembles, such as XGBoost, LightGBM, and CatBoost, still outperform the deep learning models on most data sets that we considered and come with the additional advantage of significantly less training time. Even though it has been six years since the XGBoost publication [53] and over twenty years since the publishing of original gradient boosting paper [102], we can state that despite much research effort in deep learning, the state of the art for tabular data remains largely unchanged. However, we observed that for very large data sets, approaches based on deep learning may still be able to achieve competitive performance and even outperform classical models. In summary, we think that a fundamental reorientation of the domain may be necessary. For now, the question of whether the use of current deep learning techniques is beneficial for tabular

data can generally be answered in the negative. This applies in particular to small heterogeneous data sets that are common in applications. Hence, instead of proposing more and more complex models, we argue that a more profound understanding of the reasons for this performance gap is needed.

Unified Benchmarking. Furthermore, our results highlight the need for unified benchmarks. There is no consensus in the machine learning community on how to make a fair and efficient comparison. Shwartz-Ziv & Armon [7] show that the choice of benchmarking data sets can have a non-negligible impact on the performance assessment. While we chose common data sets with varying characteristics for our experiments, a different choice of data sets or hyperparameter such as the encoding use (e.g., use one-hot encoding) may lead to a different outcome. Because of the excessive number of data sets (in the eighteen works listed in Table II, over 100 different data sets are used), there is a necessity for a standardized benchmarking procedure, which allows to identify significant progress with respect to the state of the art. With this work, we also propose an open-source benchmark for deep learning models on tabular data. For *tabular data generation* tasks, Xu et al. [144] proposes a sound evaluation framework with artificial and real-world data sets (Sec. V-B), but researchers need to agree on common benchmarks in this subdomain as well.

Tabular Data Preprocessing. Many of the challenges for deep neural networks on tabular data are related to the heterogeneity of the data (e.g., categorical and sparse values). Therefore, some deep learning solutions transform them into a homogeneous representation more suitable to neural networks. While the additional overhead is small, such transforms can boost performance considerably and should thus be among the first strategies applied in real-world scenarios.

Architectures for Deep Learning on Tabular Data. Architecture-wise, there has been a clear trend towards transformer-based solutions (Section IV-B2) in recent years. These approaches offer multiple advantages over standard neural network architectures, for instance, learning with attention over both categorical and numerical features. Moreover, self-supervised or unsupervised pre-training that leverages unlabelled tabular data to train parts of the deep learning model is gaining popularity, not only among transformer-based approaches. Performance-wise, multiple independent evaluations demonstrate that deep neural network methods from the hybrid (Sec. IV-B1) and transformers-based (Sec. IV-B2) groups exhibit superior predictive performance compared to plain deep neural networks on various data sets [9], [55], [70], [93]. This underlines the importance of special-purpose architectures for tabular data.

Regularization Models for Tabular Data. It has also been shown that regularization reduces the hypersensitivity of deep neural network models and improves the overall performance [10], [72]. We believe that regularization is one of the crucial aspects for a more robust and accurate performance of deep neural networks on tabular data and is gaining momentum.

Deep Generative Models for Tabular Data. Powerful tabular data generation is essential for the development of high-quality models, particularly in a privacy context. With suitable data generators at hand, developers can use large, synthetic, and yet realistic data sets to develop better models, while not being subject to privacy concerns [148]. Unfortunately, the generation task is as hard as inference in predictive models, so progress in both areas will likely go hand in hand.

Interpretable Deep Learning Models for Tabular Data. Interpretability is undoubtedly desirable, particularly for tabular data models frequently applied to personal data, e.g., in healthcare and finance. An increasing number of approaches offer it out-of-the-box but most current deep neural network models are still mainly concerned with the optimization of a chosen error metric. Therefore, extending existing open-source libraries (see [157], [170]) aimed at interpreting black-box models helps advance the field. Moreover, interpretable deep tabular learning is essential for understanding model decisions and results, especially for life-critical applications. However, much of the state-of-the-art recourse literature does not offer easy support of heterogeneous tabular data and lacks metrics to evaluate the quality of heterogeneous data recourse. Finally, model explanations can also be used to identify and mitigate potential bias or eliminate unfair discrimination against certain groups [208].

Learning From Evolving Data Streams. Many modern applications are subject to continuously evolving data streams, e.g., social media, online retail, or healthcare. Streaming data are usually heterogeneous and potentially unlimited. Therefore, observations must be processed in a single pass and cannot be stored. Indeed, online learning models can only access a fraction of the data at each time step. Furthermore, they have to deal with limited resources and shifting data distributions (i.e., concept drift). Hence, hyperparameter optimization and model selection, as typically involved in deep learning, are usually not feasible in a data stream. For this reason, despite the success of deep learning in other domains, less complex methods such as incremental decision trees [209], [210] are often preferred in online learning applications.

B. Open Research Questions

Several open problems need to be addressed in future research. In this section, we will list those we deem fundamental to the domain.

Information-theoretic Analysis of Encodings. Encoding methods are highly popular when dealing with tabular data. However, the majority of data preprocessing approaches for deep neural networks are lossy in terms of information content. Therefore, it is challenging to achieve an efficient, almost lossless transformation of heterogeneous tabular data into homogeneous data. Nevertheless, the information-theoretic view on these transformations remains to be investigated in detail and could shed light on the underlying mechanisms.

Computational Efficiency in Hybrid Models. The work by Shwartz-Ziv & Armon [7] suggests that the combination of a gradient boosting decision tree and deep neural networks may improve the predictive performance of a machine learning system. However, it also leads to growing complexity. Training or inference times, which far exceed those of classical machine learning approaches, are a recurring problem when developing

hybrid models. We conclude that the integration of state-of-the-art approaches from classical machine learning and deep learning has not been conclusively resolved yet and future work should be conducted on how to mitigate the trade-off between predictive performance and computational complexity.

Specialized Regularizations. We applaud recent research on regularization methods, in which we see a promising direction that necessitates further exploration. Whether context- and architecture-specific regularizations for tabular data can be found remains an open question. However, a recent work [211] indicates that regularization techniques for deep neural networks such as weight decay and data augmentation produce an unfair model across classes. Additionally, it is relevant to explore the theoretical constraints that govern the success of regularization on tabular data more profoundly.

Novel Processes for Tabular Data Generation. For tabular data generation, modified Generative Adversarial Networks and Variational Autoencoders are prevalent. However, the modelling of dependencies and categorical distributions remains the key challenge. Novel architectures in this area, such as diffusion models, have not been adapted to the domain of tabular data. Furthermore, the definition of an entirely new generative process particularly focused on tabular data might be worth investigating.

Interpretability. Going forward, counterfactual explanations for deep tabular learning can be used to improve the perceived fairness in human-AI interaction scenarios and to enable personalized decision-making [190]. However, the heterogeneity of tabular data poses problems for counterfactual explanation methods to be reliably deployed in practice. Devising techniques aimed at effectively handling heterogeneous tabular data in the presence of feasibility constraints is still an unsolved task [157].

Transfer of Deep Learning Methods to Data Streams. Recent work shows that some of the limitations of neural networks in an evolving data stream can be overcome [25], [212]. Conversely, changes in the parameters of a neural network may be effectively used to weigh the importance of input features over time [213] or to detect concept drift [214]. Accordingly, we argue that deep learning for streaming data – in particular strategies for dealing with evolving and heterogeneous tabular data – should receive more attention in the future.

Transfer Learning for Tabular Data. Reusing knowledge gained solving one problem and applying it to a different task is the research problem addressed by transfer learning. While transfer learning is successfully used in computer vision and natural language processing applications [215], there are no efficient and generally accepted ways to do transfer learning for tabular data. Hence, a general research question can be how to share knowledge between multiple (related) tabular data sets efficiently.

Data Augmentation for Tabular Data. Data augmentation has proven highly effective to prevent overfitting, especially in computer vision [216]. While some data augmentation techniques for tabular data exist, e.g., SMOTE-NC [217], simple models fail to capture the dependency structure of the data. Therefore, generating additional samples in a continuous latent space is a promising direction. This was investigated by Darabi & Elor [44] for minority oversampling. Nevertheless, the reported improvements are only marginal. Thus, future work is required to find simple, yet effective random transformations to enhance tabular training sets.

Self-supervised Learning. Large-scale labelled data are usually required to train deep neural networks; however, the data labelling is an expensive task. To avoid this expensive step, self-supervised methods propose to learn general feature representations from available unlabelled data. These methods have also shown astonishing results in computer vision and natural language processing [218], [219]. Only a few recent works in this direction [88], [89], [220] deal with heterogeneous data. Hence, novel self-supervised learning approaches dedicated to tabular data might be worth investigating.

IX. Conclusion

This survey is the first work to systematically explore deep neural network approaches for heterogeneous tabular data. In this context, we highlighted the main challenges and research advances in modelling, generating, and explaining tabular data. We introduced a unified taxonomy that categorizes deep learning approaches for tabular data into three branches: data transformation methods, specialized architectures, and regularization models. We believe our taxonomy will help catalogue future research and better understand and address the remaining challenges in applying deep learning to tabular data. We hope it will help researchers and practitioners to find the most appropriate strategies and methods for their applications.

Additionally, we also conducted an unbiased evaluation of the state-of-the-art deep learning approaches on multiple real-world data sets. Deep neural network-based methods for heterogeneous tabular data are still inferior to machine learning methods based on decision tree ensembles for small and medium-sized data sets (less than $\sim$1M samples). Only for a very large data set mainly consisting of continuous and numerical variables, the deep learning model SAINT outperformed these classical approaches. Furthermore, we assessed explanation properties of deep learning models with the self-attention mechanism. Although the TabNet model shows promising explanation explanatory capabilities, inconsistencies between the explanations remain an open issue.

Due to the importance of tabular data to industry and academia, new ideas in this area are in high demand and can have a significant impact. With this review, we hope to provide interested readers with the references and insights they need to address open challenges and effectively advance the field.

References

[1] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural networks*, vol. 61, pp. 85–117, 2015.
[2] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.
[3] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[5] S. O. Arik and T. Pfister, "TabNet: Attentive interpretable tabular learning," *arxiv:1908.07442*, 2019.

[6] S. Popov, S. Morozov, and A. Babenko, "Neural oblivious decision ensembles for deep learning on tabular data," *arxiv:1909.06312*, 2019.

[7] R. Shwartz-Ziv and A. Armon, "Tabular Data: Deep Learning is Not All You Need," *arXiv preprint arXiv:2106.03253*, 2021.

[8] S. Elsayed, D. Thyssens, A. Rashed, H. S. Jomaa, and L. Schmidt-Thieme, "Do we really need deep learning models for time series forecasting?" *arXiv preprint arXiv:2101.02118*, 2021.

[9] G. Somepalli, M. Goldblum, A. Schwarzschild, C. B. Bruss, and T. Goldstein, "SAINT: Improved neural networks for tabular data via row attention and contrastive pre-training," *arXiv preprint arXiv:2106.01342*, 2021.

[10] A. Kadra, M. Lindauer, F. Hutter, and J. Grabocka, "Well-tuned Simple Nets Excel on Tabular Datasets," in *Advances in Neural Information Processing Systems*, 2021.

[11] D. Ulmer, L. Meijerink, and G. Cinà, "Trust issues: Uncertainty estimation does not enable reliable ood detection on medical tabular data," in *Machine Learning for Health*. PMLR, 2020, pp. 341–354.

[12] S. Somani, A. J. Russak, F. Richter, S. Zhao, A. Vaid, F. Chaudhry, J. K. De Freitas, N. Naik, R. Miotto, G. N. Nadkarni *et al.*, "Deep learning and the electrocardiogram: review of the current state-of-the-art," *EP Europace*, 2021.

[13] V. Borisov, E. Kasneci, and G. Kasneci, "Robust cognitive load detection from wrist-band sensors," *Computers in Human Behavior Reports*, vol. 4, p. 100116, 2021.

[14] J. M. Clements, D. Xu, N. Yousefi, and D. Efimov, "Sequential deep learning for credit risk monitoring with tabular financial data," *arXiv preprint arXiv:2012.15330*, 2020.

[15] H. Guo, R. Tang, Y. Ye, Z. Li, and X. He, "Deepfm: a factorization-machine based neural network for ctr prediction," *arXiv preprint arXiv:1703.04247*, 2017.

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

[17] M. Ahmed, H. Afzal, A. Majeed, and B. Khan, "A survey of evolution in predictive models and impacting factors in customer churn," *Advances in Data Science and Adaptive Analysis*, vol. 9, no. 03, p. 1750007, 2017.

[18] Q. Tang, G. Xia, X. Zhang, and F. Long, "A customer churn prediction model based on xgboost and mlp," in *2020 International Conference on Computer Engineering and Application (ICCEA)*. IEEE, 2020, pp. 608–612.

[19] A. L. Buczak and E. Guven, "A survey of data mining and machine learning methods for cyber security intrusion detection," *IEEE Communications surveys & tutorials*, vol. 18, no. 2, pp. 1153–1176, 2015.

[20] F. Cartella, O. Anunciação, Y. Funabiki, D. Yamaguchi, T. Akishita, and O. Elshocht, "Adversarial attacks for tabular data: application to fraud detection and imbalanced data," *CEUR Workshop Proceedings*, vol. 2808, 2021.

[21] B. Liu, M. Ding, S. Shaham, W. Rahayu, F. Farokhi, and Z. Lin, "When machine learning meets privacy: A survey and outlook," *ACM Computing Surveys (CSUR)*, vol. 54, no. 2, pp. 1–36, 2021.

[22] C. J. Urban and K. M. Gates, "Deep learning: A primer for psychologists." *Psychological Methods*, 2021.

[23] M. Shoman, A. Aboah, and Y. Adu-Gyamfi, "Deep learning framework for predicting bus delays on multiple routes using heterogenous datasets," *Journal of Big Data Analytics in Transportation*, vol. 2, no. 3, pp. 275–290, 2020.

[24] G. Pang, C. Shen, L. Cao, and A. V. D. Hengel, "Deep learning for anomaly detection: A review," *ACM Computing Surveys (CSUR)*, vol. 54, no. 2, pp. 1–38, 2021.

[25] D. Sahoo, Q. Pham, J. Lu, and S. C. Hoi, "Online deep learning: Learning deep neural networks on the fly," *arXiv preprint arXiv:1711.03705*, 2017.

[26] X. He, K. Zhao, and X. Chu, "AutoML: A survey of the state-of-the-art," *Knowledge-Based Systems*, vol. 212, p. 106622, 2021.

[27] M. Artzi, E. Redmard, O. Tzemach, J. Zeltser, O. Gropper, J. Roth, B. Shofty, D. A. Kozyrev, S. Constantini, and L. Ben-Sira, "Classification of pediatric posterior fossa tumors using convolutional neural network and tabular data," *IEEE Access*, vol. 9, pp. 91 966–91 973, 2021.

[28] X. Shi, J. Mueller, N. Erickson, M. Li, and A. Smola, "Multimodal AutoML on structured tables with text fields," in *8th ICML Workshop on Automated Machine Learning (AutoML)*, 2021.

[29] R. Fakoor, J. W. Mueller, N. Erickson, P. Chaudhari, and A. J. Smola, "Fast, accurate, and simple models for tabular data via augmented distillation," *Advances in Neural Information Processing Systems*, vol. 34, 2020.

[30] P. Gijsbers, E. LeDell, J. Thomas, S. Poirier, B. Bischl, and J. Vanschoren, "An open source AutoML benchmark," *arXiv preprint arXiv:1907.00909*, 2019.

[31] P. Yin, G. Neubig, W.-t. Yih, and S. Riedel, "TaBERT: Pretraining for joint understanding of textual and tabular data," *arxiv:2005.08314*, 2020.

[32] Z. Wang, Q. She, and T. E. Ward, "Generative adversarial networks in computer vision: A survey and taxonomy," *arXiv preprint arXiv:1906.01529*, 2019.

[33] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 2, pp. 423–443, 2018.

[34] D. Lichtenwalter, P. Burggräf, J. Wagner, and T. Weißer, "Deep multimodal learning for manufacturing problem solving," *Procedia CIRP*, vol. 99, pp. 615–620, 2021.

[35] S. Pölsterl, T. N. Wolf, and C. Wachinger, "Combining 3d image and tabular data via the dynamic affine feature map transform," *arXiv preprint arXiv:2107.05990*, 2021.

[36] D. d. B. Soares, F. Andrieux, B. Hell, J. Lenhardt, J. Badosa, S. Gavoille, S. Gaiffas, and E. Bacry, "Predicting the solar potential of rooftops using image segmentation and structured data," *arXiv preprint arXiv:2106.15268*, 2021.

[37] D. Medvedev and A. D'yakonov, "New properties of the data distillation method when working with tabular data," *arXiv preprint arXiv:2010.09839*, 2020.

[38] J. Li, Y. Li, X. Xiang, S.-T. Xia, S. Dong, and Y. Cai, "Tnt: An interpretable tree-network-tree learning framework using knowledge distillation," *Entropy*, vol. 22, no. 11, p. 1203, 2020.

[39] D. Roschewitz, M.-A. Hartley, L. Corinzia, and M. Jaggi, "Ifedavg: Interpretable data-interoperability for federated learning," *arXiv preprint arXiv:2107.06580*, 2021.

[40] A. Sánchez-Morales, J.-L. Sancho-Gómez, J.-A. Martínez-García, and A. R. Figueiras-Vidal, "Improving deep learning performance with missing values via deletion and compensation," *Neural Computing and Applications*, vol. 32, no. 17, pp. 13 233–13 244, 2020.

[41] L. Gondara and K. Wang, "Mida: Multiple imputation using denoising autoencoders," in *Pacific-Asia conference on knowledge discovery and data mining*. Springer, 2018, pp. 260–272.

[42] R. D. Camino, C. Hammerschmidt *et al.*, "Working with deep generative models and tabular data imputation," *ICML 2020 Artemiss Workshop*, 2020.

[43] J. Engelmann and S. Lessmann, "Conditional wasserstein gan-based oversampling of tabular data for imbalanced learning," *Expert Systems with Applications*, vol. 174, p. 114582, 2021.

[44] S. Darabi and Y. Elor, "Synthesising multi-modal minority samples for tabular data," *arXiv preprint arXiv:2105.08204*, 2021.

[45] S. Kamthe, S. Assefa, and M. Deisenroth, "Copula flows for synthetic data generation," *arXiv preprint arXiv:2101.00598*, 2021.

[46] E. Choi, S. Biswal, B. Malin, J. Duke, W. F. Stewart, and J. Sun, "Generating Multi-label Discrete Patient Records using Generative Adversarial Networks," *Machine learning for healthcare conference*, pp. 286–305, 2017.

[47] C. OAG, "Ccpa regulations: Final regulation text." *Office of the Attorney General, California Department of Justice*, 2021.

[48] GDPR, "Regulation (eu) 2016/679 of the european parliament and of the council," *Official Journal of the European Union*, 2016. [Online]. Available: http://www.privacyregulation.eu/en/13.htm

[49] P. Voigt and A. Von dem Bussche, "The eu general data protection regulation (gdpr)," *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, vol. 10, p. 3152676, 2017.

[50] M. Sahakyan, Z. Aung, and T. Rahwan, "Explainable artificial intelligence for tabular data: A survey," *IEEE Access*, 2021.

[51] B. I. Grisci, M. J. Krause, and M. Dorn, "Relevance aggregation for neural networks interpretability and knowledge discovery on tabular data," *Information Sciences*, vol. 559, pp. 111–129, 2021.

[52] U. Bhatt, A. Xiang, S. Sharma, A. Weller, A. Taly, Y. Jia, J. Ghosh, R. Puri, J. M. Moura, and P. Eckersley, "Explainable machine learning in deployment," in *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 2020, pp. 648–657.

[53] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2016, pp. 785–794.

[54] J. T. Hancock and T. M. Khoshgoftaar, "Survey on categorical data for neural networks," *Journal of Big Data*, vol. 7, pp. 1–41, 2020.

[55] Y. Gorishniy, I. Rubachev, V. Khrulkov, and A. Babenko, "Revisiting deep learning models for tabular data," *arXiv preprint arXiv:2106.11959*, 2021.

[56] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[57] L. Katzir, G. Elidan, and R. El-Yaniv, "Net-DNF: Effective deep modeling of tabular data," in *International Conference on Learning Representations*, 2021.

[58] R. U. David M. Lane, *Introduction to Statistics*. David Lane, 2003.

[59] M. Ryan, *Deep learning with structured data*. Simon and Schuster, 2020.

[60] M. W. Cvitkovic *et al.*, "Deep learning in unconventional domains," Ph.D. dissertation, California Institute of Technology, 2020.

[61] D. Dua and C. Graff, "UCI machine learning repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml

[62] A. J. Miles, "The sunstroke epidemic of cincinnati, ohio, during the summer of 1881," *Public health papers and reports*, vol. 7, p. 293, 1881.

[63] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of eugenics*, vol. 7, no. 2, pp. 179–188, 1936.

[64] D. A. Jdanov, D. Jasilionis, V. M. Shkolnikov, and M. Barbieri, "Human mortality database," *Encyclopedia of gerontology and population aging/editors Danan Gu, Matthew E. Dupre. Cham: Springer International Publishing, 2020*, 2019.

[65] E. Fix, *Discriminatory analysis: nonparametric discrimination, consistency properties*. USAF school of Aviation Medicine, 1951.

[66] C. L. Giles, C. B. Miller, D. Chen, H.-H. Chen, G.-Z. Sun, and Y.-C. Lee, "Learning and extracting finite state automata with second-order recurrent neural networks," *Neural Computation*, vol. 4, no. 3, pp. 393–405, 1992.

[67] B. G. Horne and C. L. Giles, "An experimental comparison of recurrent neural networks," *Advances in neural information processing systems*, pp. 697–704, 1995.

[68] L. Willenborg and T. De Waal, *Statistical disclosure control in practice*. Springer Science & Business Media, 1996, vol. 111.

[69] M. Richardson, E. Dominowska, and R. Ragno, "Predicting clicks: estimating the click-through rate for new ads," in *Proceedings of the 16th international conference on World Wide Web*, 2007, pp. 521–530.

[70] G. Ke, Z. Xu, J. Zhang, J. Bian, and T.-Y. 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*, 2019, pp. 384–394.

[71] Z. Wang, Q. She, and J. Zhang, "Masknet: Introducing feature-wise multiplication to ctr ranking models by instance-guided mask," *arXiv:2102.07619*, 2021.

[72] I. Shavitt and E. Segal, "Regularization learning networks: deep learning for tabular datasets," in *Advances in Neural Information Processing Systems*, 2018, pp. 1379–1389.

[73] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *arXiv: 2005.14165*, 2020.

[74] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.

[75] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *arXiv preprint arXiv:2101.01169*, 2021.

[76] A. F. Karr, A. P. Sanil, and D. L. Banks, "Data quality: A statistical perspective," *Statistical Methodology*, vol. 3, no. 2, pp. 137–173, 2006.

[77] L. Xu and K. Veeramachaneni, "Synthesizing Tabular Data using Generative Adversarial Networks," *arXiv preprint arXiv:1811.11264*, 2018.

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

[79] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "Catboost: unbiased boosting with categorical features," in *Advances in neural information processing systems*, 2018, pp. 6638–6648.

[80] Y. Zhu, T. Brettin, F. Xia, A. Partin, M. Shukla, H. Yoo, Y. A. Evrard, J. H. Doroshow, and R. L. Stevens, "Converting tabular data into images for deep learning with convolutional neural networks," *Scientific Reports*, vol. 11, no. 1, pp. 1–11, 2021.

[81] N. Rahaman, A. Baratin, D. Arpit, F. Draxler, M. Lin, F. Hamprecht, Y. Bengio, and A. Courville, "On the spectral bias of neural networks," in *International Conference on Machine Learning*. PMLR, 2019, pp. 5301–5310.

[82] B. R. Mitchell *et al.*, "The spatial inductive bias of deep learning," Ph.D. dissertation, Johns Hopkins University, 2017.

[83] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[84] Y. Gorishniy, I. Rubachev, and A. Babenko, "On embeddings for numerical features in tabular deep learning," *arXiv preprint arXiv:2203.05556*, 2022.

[85] E. Fitkov-Norris, S. Vahid, and C. Hand, "Evaluating the impact of categorical data encoding and scaling on neural network classification performance: the case of repeat consumption of identical cultural goods," in *International Conference on Engineering Applications of Neural Networks*. Springer, 2012, pp. 343–352.

[86] D. Baylor, E. Breck, H.-T. Cheng, N. Fiedel, C. Y. Foo, Z. Haque, S. Haykal, M. Ispir, V. Jain, L. Koc *et al.*, "TFX: A tensorflow-based production-scale machine learning platform," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017, pp. 1387–1395.

[87] B. Sun, L. Yang, W. Zhang, M. Lin, P. Dong, C. Young, and J. Dong, "Supertml: Two-dimensional word embedding for the precognition on structured tabular data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.

[88] J. Yoon, Y. Zhang, J. Jordon, and M. van der Schaar, "Vime: Extending the success of self-and semi-supervised learning to tabular domaindim," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[89] D. Bahri, H. Jiang, Y. Tay, and D. Metzler, "SCARF: Self-supervised contrastive learning using random feature corruption," *arXiv preprint arXiv:2106.15147*, 2021.

[90] H.-T. Cheng, L. Koc, J. Harmsen, T. Shaked, T. Chandra, H. Aradhye, G. Anderson, G. Corrado, W. Chai, M. Ispir *et al.*, "Wide & deep learning for recommender systems," in *Proceedings of the 1st workshop on deep learning for recommender systems*, 2016, pp. 7–10.

[91] N. Frosst and G. Hinton, "Distilling a neural network into a soft decision tree," *arXiv preprint arXiv:1711.09784*, 2017.

[92] J. Lian, X. Zhou, F. Zhang, Z. Chen, X. Xie, and G. Sun, "xdeepfm: Combining explicit and implicit feature interactions for recommender systems," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018, pp. 1754–1763.

[93] G. Ke, J. Zhang, Z. Xu, J. Bian, and T.-Y. Liu, "TabNN: A universal neural network solution for tabular data," 2018.

[94] Y. Luo, H. Zhou, W.-W. Tu, Y. Chen, W. Dai, and Q. 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*, 2020, pp. 2317–2326.

[95] Z. Liu, Q. Liu, H. Zhang, and Y. Chen, "Dnn2lr: Interpretation-inspired feature crossing for real-world tabular data," *arXiv preprint arXiv:2008.09775*, 2020.

[96] S. Ivanov and L. Prokhorenkova, "Boost then Convolve: Gradient Boosting Meets Graph Neural Networks," in *Interational Conference on Learning Representations*, 2021.

[97] H. Luo, F. Cheng, H. Yu, and Y. Yi, "SDTR: Soft Decision Tree Regressor for Tabular Data," *IEEE Access*, vol. 9, pp. 55 999–56 011, 2021.

[98] X. Huang, A. Khetan, M. Cvitkovic, and Z. Karnin, "TabTransformer: Tabular Data Modeling Using Contextual Embeddings," *arxiv:2012.06678*, 2020.

[99] S. Cai, K. Zheng, G. Chen, H. Jagadish, B. C. Ooi, and M. Zhang, "Arm-net: Adaptive relation modeling network for structured data," in *Proceedings of the 2021 International Conference on Management of Data*, 2021, pp. 207–220.

[100] J. Kossen, N. Band, C. Lyle, A. Gomez, T. Rainforth, and Y. Gal, "Self-attention between datapoints: Going beyond individual input-output pairs in deep learning," in *Advances in Neural Information Processing Systems*, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., 2021.

[101] D. Micci-Barreca, "A preprocessing scheme for high-cardinality categorical attributes in classification and prediction problems," *SIGKDD Explor.*, vol. 3, pp. 27–32, 2001.

[102] J. H. Friedman, "Stochastic gradient boosting," *Computational statistics & data analysis*, vol. 38, no. 4, pp. 367–378, 2002.

[103] P. Langley and S. Sage, "Oblivious decision trees and abstract cases," in *Working notes of the AAAI-94 workshop on case-based reasoning*. Seattle, WA, 1994, pp. 113–117.

[104] B. Peters, V. Niculae, and A. F. Martins, "Sparse sequence-to-sequence models," *arxiv:1905.05702*, 2019.

[105] Y. LeCun and C. Cortes, "MNIST handwritten digit database," 2010. [Online]. Available: http://yann.lecun.com/exdb/mnist/

[106] C. Guo and F. Berkhahn, "Entity embeddings of categorical variables," *arXiv preprint arXiv:1604.06737*, 2016.

[107] S. Rendle, "Factorization machines," in *2010 IEEE International conference on data mining*. IEEE, 2010, pp. 995–1000.

[108] F. Moosmann, B. Triggs, and F. Jurie, "Fast discriminative visual codebooks using randomized clustering forests," in *Twentieth Annual Conference on Neural Information Processing Systems (NIPS'06)*. MIT Press, 2006, pp. 985–992.

[109] X. He, J. Pan, O. Jin, T. Xu, B. Liu, T. Xu, Y. Shi, A. Atallah, R. Herbrich, S. Bowers *et al.*, "Practical lessons from predicting clicks on ads at facebook," in *Proceedings of the Eighth International Workshop on Data Mining for Online Advertising*, 2014, pp. 1–9.

[110] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," *IEEE transactions on neural networks*, vol. 20, no. 1, pp. 61–80, 2008.

[111] C. Wang, M. Li, and A. J. Smola, "Language models with transformers," *arXiv preprint arXiv:1904.09408*, 2019.

[112] A. F. T. Martins and R. F. Astudillo, "From softmax to sparsemax: A sparse model of attention and multi-label classification," *arxiv:1602.02068*, 2016.

[113] G. Van Rossum and F. L. Drake Jr, *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam, 1995.

[114] M. Joseph, "Pytorch tabular: A framework for deep learning with tabular data," *arXiv preprint arXiv:2104.13638*, 2021.

[115] S. Boughorbel, F. Jarray, and A. Kadri, "Fairness in tabnet model by disentangled representation for the prediction of hospital no-show," *arXiv preprint arXiv:2103.04048*, 2021.

[116] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Computing Surveys (CSUR)*, vol. 54, no. 6, pp. 1–35, 2021.

[117] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "Cutmix: Regularization strategy to train strong classifiers with localizable features," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6023–6032.

[118] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

[119] V. Borisov, J. Haug, and G. Kasneci, "CancelOut: A layer for feature selection in deep neural networks," in *International Conference on Artificial Neural Networks*. Springer, 2019, pp. 72–83.

[120] G. Valdes, W. Arbelo, Y. Interian, and J. H. Friedman, "Lockout: Sparse regularization of neural networks," *arXiv preprint arXiv:2107.07160*, 2021.

[121] J. Fiedler, "Simple modifications to improve tabular neural networks," *arXiv preprint arXiv:2108.03214*, 2021.

[122] K. Lounici, K. Meziani, and B. Riu, "Muddling label regularization: Deep learning for tabular datasets," *arXiv preprint arXiv:2106.04462*, 2021.

[123] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *NeurIPS*, 2017.

[124] H. Chen, S. Jajodia, J. Liu, N. Park, V. Sokolov, and V. Subrahmanian, "Faketables: Using gans to generate functional dependency preserving tables with bounded real data." in *IJCAI*, 2019, pp. 2074–2080.

[125] M. Quintana and C. Miller, "Towards class-balancing human comfort datasets with gans," in *Proceedings of the 6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*, 2019, pp. 391–392.

[126] A. Koivu, M. Sairanen, A. Airola, and T. Pahikkala, "Synthetic minority oversampling of vital statistics data with generative adversarial networks," *Journal of the American Medical Informatics Association*, vol. 27, no. 11, pp. 1667–1674, 2020.

[127] J. Fan, J. Chen, T. Liu, Y. Shen, G. Li, and X. Du, "Relational data synthesis using generative adversarial networks: A design space exploration," *Proc. VLDB Endow.*, vol. 13, no. 12, p. 1962–1975, Jul. 2020.

[128] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8110–8119.

[129] K. Lin, D. Li, X. He, Z. Zhang, and M.-T. Sun, "Adversarial ranking for language generation," *Advances in Neural Information Processing Systems*, 2017.

[130] S. Subramanian, S. Rajeswar, F. Dutil, C. Pal, and A. Courville, "Adversarial generation of natural language," in *Proceedings of the 2nd Workshop on Representation Learning for NLP*, 2017, pp. 241–251.

[131] N. Patki, R. Wedge, and K. Veeramachaneni, "The synthetic data vault," in *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, 2016, pp. 399–410.

[132] Z. Li, Y. Zhao, and J. Fu, "Sync: A copula based framework for generating synthetic data from aggregated sources," in *2020 International Conference on Data Mining Workshops (ICDMW)*. IEEE, 2020, pp. 571–578.

[133] J. Zhang, G. Cormode, C. M. Procopiuc, D. Srivastava, and X. Xiao, "Privbayes: Private data release via bayesian networks," *ACM Transactions on Database Systems (TODS)*, vol. 42, no. 4, pp. 1–41, 2017.

[134] C. Chow and C. Liu, "Approximating discrete probability distributions with dependence trees," *IEEE transactions on Information Theory*, vol. 14, no. 3, pp. 462–467, 1968.

[135] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *arXiv preprint arXiv:1406.2661*, 2014.

[136] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *4th International Conference on Learning Representations, ICLR 2016*, 2016.

[137] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *International conference on machine learning*. PMLR, 2017, pp. 214–223.

[138] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved training of wasserstein gans," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 5769–5779.

[139] M. G. Bellemare, I. Danihelka, W. Dabney, S. Mohamed, B. Lakshminarayanan, S. Hoyer, and R. Munos, "The cramer distance as a solution to biased wasserstein gradients," 2017.

[140] R. D. Hjelm, A. P. Jacob, T. Che, A. Trischler, K. Cho, and Y. Bengio, "Boundary-seeking generative adversarial networks," *International Conference on Learning Representations*, 2018.

[141] A. Srivastava, L. Valkov, C. Russell, M. U. Gutmann, and C. Sutton, "Veegan: Reducing mode collapse in gans using implicit variational learning," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 3310–3320.

[142] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings*, no. Ml, pp. 1–14, 2014.

[143] C. Ma, S. Tschiatschek, R. Turner, J. M. Hernández-Lobato, and C. Zhang, "Vaem: a deep generative model for heterogeneous mixed type data," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[144] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, "Modeling tabular data using conditional GAN," in *Advances in Neural Information Processing Systems*, vol. 33, 2019.

[145] N. Park, M. Mohammadi, K. Gorde, S. Jajodia, H. Park, and Y. Kim, "Data synthesis based on generative adversarial networks," *Proceedings of the VLDB Endowment*, vol. 11, no. 10, pp. 1071–1083, 2018.

[146] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with gumbel-softmax," in *International Conference on Learning Representations*, 2017.

[147] B. Wen, L. O. Colon, K. Subbalakshmi, and R. Chandramouli, "Causal-TGAN: Generating tabular data using causal generative adversarial networks," *arXiv preprint arXiv:2104.10680*, 2021.

[148] J. Jordon, J. Yoon, and M. Van Der Schaar, "Pate-gan: Generating synthetic data with differential privacy guarantees," in *International conference on learning representations*, 2018.

[149] A. Mottini, A. Lheritier, and R. Acuna-Agost, "Airline Passenger Name Record Generation using Generative Adversarial Networks," *arXiv preprint arXiv:1807.06657*, 2018.

[150] R. Camino, C. Hammerschmidt, and R. State, "Generating multi-categorical samples with generative adversarial networks," *ICML workshop on Theoretical Foundations and Applications of Deep Generative Models*, 2018.

[151] M. K. Baowaly, C. C. Lin, C. L. Liu, and K. T. Chen, "Synthesizing electronic health records using improved generative adversarial networks," *Journal of the American Medical Informatics Association*, vol. 26, no. 3, pp. 228–241, 2019.

[152] L. V. H. Vardhan and S. Kok, "Generating privacy-preserving synthetic tabular data using oblivious variational autoencoders," in *Proceedings of the Workshop on Economics of Privacy and Data Labor at the 37 th International Conference on Machine Learning*, 2020.

[153] Z. Zhao, A. Kunar, H. Van der Scheer, R. Birke, and L. Y. Chen, "Ctab-gan: Effective table data synthesizing," *arXiv preprint arXiv:2102.08369*, 2021.

[154] F. J. Massey Jr, "The kolmogorov-smirnov test for goodness of fit," *Journal of the American statistical Association*, vol. 46, no. 253, pp. 68–78, 1951.

[155] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM Comput. Surv.*, vol. 51, no. 5, 2018.

[156] K. Gade, S. C. Geyik, K. Kenthapadi, V. Mithal, and A. Taly, "Explainable ai in industry," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2019.

[157] M. Pawelczyk, S. Bielawski, J. Van den Heuvel, T. Richter, and G. Kasneci, "Carla: A python library to benchmark algorithmic recourse and counterfactual explanation algorithms," in *Advances in Neural Information Processing Systems (NeurIPS) Benchmark and Datasets Track*, 2021.

[158] Y. Lou, R. Caruana, and J. Gehrke, "Intelligible models for classification and regression," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2012.

[159] D. Alvarez-Melis and T. S. Jaakkola, "Towards robust interpretability with self-explaining neural networks," *NeurIPS*, 2018.

[160] D. Wang, Q. Yang, A. Abdul, and B. Y. Lim, "Designing theory-driven user-centric explainable ai," in *CHI*, 2019.

[161] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PloS one*, vol. 10, no. 7, p. e0130140, 2015.

[162] G. Montavon, A. Binder, S. Lapuschkin, W. Samek, and K.-R. Müller, "Layer-wise relevance propagation: an overview," *Explainable AI: interpreting, explaining and visualizing deep learning*, pp. 193–209, 2019.

[163] G. Kasneci and T. Gottron, "Licon: A linear weighting scheme for the contribution ofinput variables in deep artificial neural networks," in *CIKM*, 2016.

[164] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *International Conference on Machine Learning*. PMLR, 2017, pp. 3319–3328.

[165] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2018.

[166] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[167] M. T. Ribeiro, S. Singh, and C.Guestrin, "Anchors: High-precision model-agnostic explanations," in *AAAI*, 2018.

[168] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, "From local explanations to global understanding with explainable ai for trees," *Nature machine intelligence*, 2020.

[169] J. Haug, S. Zürn, P. El-Jiz, and G. Kasneci, "On baselines for local feature attributions," *arXiv: 2101.00905*, 2021.

[170] Y. Liu, S. Khandagale, C. White, and W. Neiswanger, "Synthetic benchmarks for scientific research in explainable machine learning," in *Advances in Neural Information Processing Systems (NeurIPS) Benchmark and Datasets Track*, 2021.

[171] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: automated decisions and the gdpr," *Harvard Journal of Law & Technology*, vol. 31, no. 2, 2018.

[172] B. Ustun, A. Spangher, and Y. Liu, "Actionable recourse in linear classification," in *(FAT*)*, 2019.

[173] C. Russell, "Efficient search for diverse coherent explanations," in *(FAT*)*, 2019.

[174] K. Rawal and H. Lakkaraju, "Beyond individualized recourse: Interpretable and interactive summaries of actionable recourses," in *NeurIPS*, 2020.

[175] A.-H. Karimi, G. Barthe, B. Balle, and I. Valera, "Model-agnostic counterfactual explanations for consequential decisions," in *AISTATS*, 2020.

[176] A. Dhurandhar, P.-Y. Chen, R. Luss, C.-C. Tu, P. Ting, K. Shanmugam, and P. Das, "Explanations based on the missing: Towards contrastive explanations with pertinent negatives," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[177] B. Mittelstadt, C. Russell, and S. Wachter, "Explaining explanations in ai," in *Proceedings of the conference on fairness, accountability, and transparency*, 2019.

[178] R. K. Mothilal, A. Sharma, and C. Tan, "Explaining machine learning classifiers through diverse counterfactual explanations," in *FAT**, 2020.

[179] M. Pawelczyk, K. Broelemann, and G. Kasneci, "Learning model-agnostic counterfactual explanations for tabular data," in *The Web Conference 2020 (WWW)*. ACM, 2020.

[180] M. Downs, J. L. Chu, Y. Yacoby, F. Doshi-Velez, and W. Pan, "Cruds: Counterfactual recourse using disentangled subspaces," *ICML Workshop on Human Interpretability in Machine Learning (WHI 2020)*, 2020.

[181] S. Joshi, O. Koyejo, W. Vijitbenjaronk, B. Kim, and J. Ghosh, "Towards realistic individual recourse and actionable explanations in black-box decision making systems," *arXiv preprint arXiv:1907.09615*, 2019.

[182] D. Mahajan, C. Tan, and A. Sharma, "Preserving causal constraints in counterfactual explanations for machine learning classifiers," *arXiv preprint arXiv:1912.03277*, 2019.

[183] M. Pawelczyk, K. Broelemann, and G. Kasneci, "On counterfactual explanations under predictive multiplicity," in *Conference on Uncertainty in Artificial Intelligence (UAI)*. PMLR, 2020.

[184] J. Antorán, U. Bhatt, T. Adel, A. Weller, and J. M. Hernández-Lobato, "Getting a clue: A method for explaining uncertainty estimates," *ICLR*, 2021.

[185] A. Nazabal, P. M. Olmos, Z. Ghahramani, and I. Valera, "Handling incomplete heterogeneous data using vaes," *Pattern Recognition*, 2020.

[186] S. Upadhyay, S. Joshi, and H. Lakkaraju, "Towards robust and reliable algorithmic recourse," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, 2021.

[187] R. Dominguez-Olmedo, A.-H. Karimi, and B. Schölkopf, "On the adversarial robustness of causal algorithmic recourse," in *International Conference on Machine Learning (ICML)*, 2022.

[188] M. Pawelczyk, T. Datta, J. van-den Heuvel, G. Kasneci, and H. Lakkaraju, "Algorithmic recourse in the face of noisy human responses," *arXiv:2203.06768*, 2022.

[189] A.-H. Karimi, G. Barthe, B. Schölkopf, and I. Valera, "A survey of algorithmic recourse: definitions, formulations, solutions, and prospects," *arXiv preprint arXiv:2010.04050*, 2020.

[190] S. Verma, J. Dickerson, and K. Hines, "Counterfactual explanations for machine learning: A review," 2020.

[191] A. Krizhevsky, "Learning multiple layers of features from tiny images," 2009.

[192] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[193] FICO, "Home equity line of credit (HELOC) dataset," https://community.fico.com/s/explainable-machine-learning-challenge, 2019 (accessed June 15, 2022).

[194] P. Baldi, P. Sadowski, and D. Whiteson, "Searching for exotic particles in high-energy physics with deep learning," *Nature communications*, vol. 5, no. 1, pp. 1–9, 2014.

[195] C. Z. Mooney, *Monte carlo simulation*. Sage, 1997, no. 116.

[196] R. K. Pace and R. Barry, "Sparse spatial autoregressions," *Statistics & Probability Letters*, vol. 33, no. 3, pp. 291–297, 1997.

[197] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and regression trees*. Routledge, 2017.

[198] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[199] K. Broelemann and G. Kasneci, "A gradient-based split criterion for highly accurate and transparent model trees," in *IJCAI*, 2019.

[200] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *The bulletin of mathematical biophysics*, vol. 5, no. 4, pp. 115–133, 1943.

[201] Y. Yamada, O. Lindenbaum, S. Negahban, and Y. Kluger, "Feature selection using stochastic gates," in *Proceedings of Machine Learning and Systems 2020*, 2020, pp. 8952–8963.

[202] R. Agarwal, N. Frosst, X. Zhang, R. Caruana, and G. E. Hinton, "Neural additive models: Interpretable machine learning with neural nets," *arXiv preprint arXiv:2004.13912*, 2020.

[203] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proceedings of the 25rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.

[204] D. Merkel, "Docker: lightweight linux containers for consistent development and deployment," *Linux journal*, vol. 2014, no. 239, p. 2, 2014.
[205] C. S. Bojer and J. P. Meldgaard, "Kaggle forecasting competitions: An overlooked learning opportunity," *International Journal of Forecasting*, vol. 37, no. 2, pp. 587–603, 2021.
[206] Y. Rong, T. Leemann, V. Borisov, G. Kasneci, and E. Kasneci, "A consistent and efficient evaluation strategy for feature attribution methods," in *International Conference on Machine Learning*. PMLR, 2022.
[207] R. Tomsett, D. Harborne, S. Chakraborty, P. Gurram, and A. Preece, "Sanity checks for saliency metrics," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, 2020, pp. 6021–6029.
[208] E. Ntoutsi, P. Fafalios, U. Gadiraju, V. Iosifidis, W. Nejdl, M.-E. Vidal, S. Ruggieri, F. Turini, S. Papadopoulos, E. Krasanakis *et al.*, "Bias in data-driven artificial intelligence systems—an introductory survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 10, no. 3, p. e1356, 2020.
[209] P. Domingos and G. Hulten, "Mining high-speed data streams," in *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2000, pp. 71–80.
[210] C. Manapragada, G. I. Webb, and M. Salehi, "Extremely fast decision tree," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018, pp. 1953–1962.
[211] R. Balestriero, L. Bottou, and Y. LeCun, "The effects of regularization and data augmentation are class dependent," *arXiv preprint arXiv:2204.03632*, 2022.
[212] P. Duda, M. Jaworski, A. Cader, and L. Wang, "On training deep neural networks using a streaming approach," *Journal of Artificial Intelligence and Soft Computing Research*, vol. 10, 2020.
[213] J. Haug, M. Pawelczyk, K. Broelemann, and G. Kasneci, "Leveraging model inherent variable importance for stable online feature selection," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 1478–1502.
[214] J. Haug and G. Kasneci, "Learning parameter distributions to detect concept drift in data streams," in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 9452–9459.
[215] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu, "A survey on deep transfer learning," in *International conference on artificial neural networks*. Springer, 2018, pp. 270–279.
[216] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, no. 1, pp. 1–48, 2019.
[217] 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.
[218] L. Jing and Y. Tian, "Self-supervised visual feature learning with deep neural networks: A survey," *IEEE transactions on pattern analysis and machine intelligence*, 2020.
[219] X. Liu, F. Zhang, Z. Hou, L. Mian, Z. Wang, J. Zhang, and J. Tang, "Self-supervised learning: Generative or contrastive," *IEEE Transactions on Knowledge and Data Engineering*, 2021.
[220] T. Ucar, E. Hajiramezanali, and L. Edwards, "Subtab: Subsetting features of tabular data for self-supervised representation learning," *Advances in Neural Information Processing Systems*, vol. 34, 2021.