A Survey of Foundation Models and Techniques for Tabular Data

www.surveyx.cn

Abstract

This survey paper explores the transformative role of foundation models in advancing tabular data analysis across various domains, including healthcare, finance, and retail. Foundation models, characterized by their pre-training on extensive datasets, enhance the efficiency and performance of machine learning applications by automating complex tasks and improving interpretability. The paper highlights the significance of tabular data in predictive modeling and addresses challenges such as high dimensionality, sparsity, and imbalanced datasets. It examines the adaptation of foundation models to tabular data, emphasizing techniques like dynamic feature weighting and contextual representation learning. The integration of pretrained models and transfer learning further enhances model performance, particularly in data-scarce scenarios. Data augmentation techniques, including generative models like TGAN, are pivotal in expanding training datasets while preserving their statistical properties. The survey also delves into representation learning, focusing on interpretability, and identifies challenges in cross-table pre-training. Future research directions include improving model scalability, enhancing privacy-preserving techniques, and integrating emerging technologies. Overall, foundation models demonstrate significant potential in processing, interpreting, and visualizing tabular data, driving innovation and necessitating continued exploration to fully harness their capabilities.

1 Introduction

1.1 Importance of Tabular Data

Tabular data is fundamental to predictive modeling across diverse industries, significantly influencing operations and decision-making processes [1]. Its structured format, characterized by rows and columns, is crucial in sectors like healthcare, finance, and retail. For instance, Electronic Health Records (EHRs) in healthcare provide essential information that enhances predictive modeling in clinical environments [2]. In finance, customer profile data, encompassing demographics and financial history, is vital for assessing financial risks [3].

In the realm of High Performance Computing (HPC), performance data is utilized to predict execution times based on configurations, optimizing job scheduling and resource utilization [4]. However, challenges such as high dimensionality and sparsity can lead to overfitting in deep learning models [5]. Moreover, imbalanced datasets necessitate advanced techniques to maintain robust model performance [6].

The complexity and inaccessibility of tabular data analysis for nonspecialists highlight the need for user-friendly tools and methodologies [7]. Additionally, standardization issues arise when tabular data is sourced from various local councils, complicating its restructuring for research purposes [8].

This survey examines interpretable representations (IRs) in explainable AI, emphasizing their role in enhancing the interpretability of black-box predictive systems [9]. This focus on interpretability is crucial for fostering understanding and trust in machine learning models applied to tabular datasets.

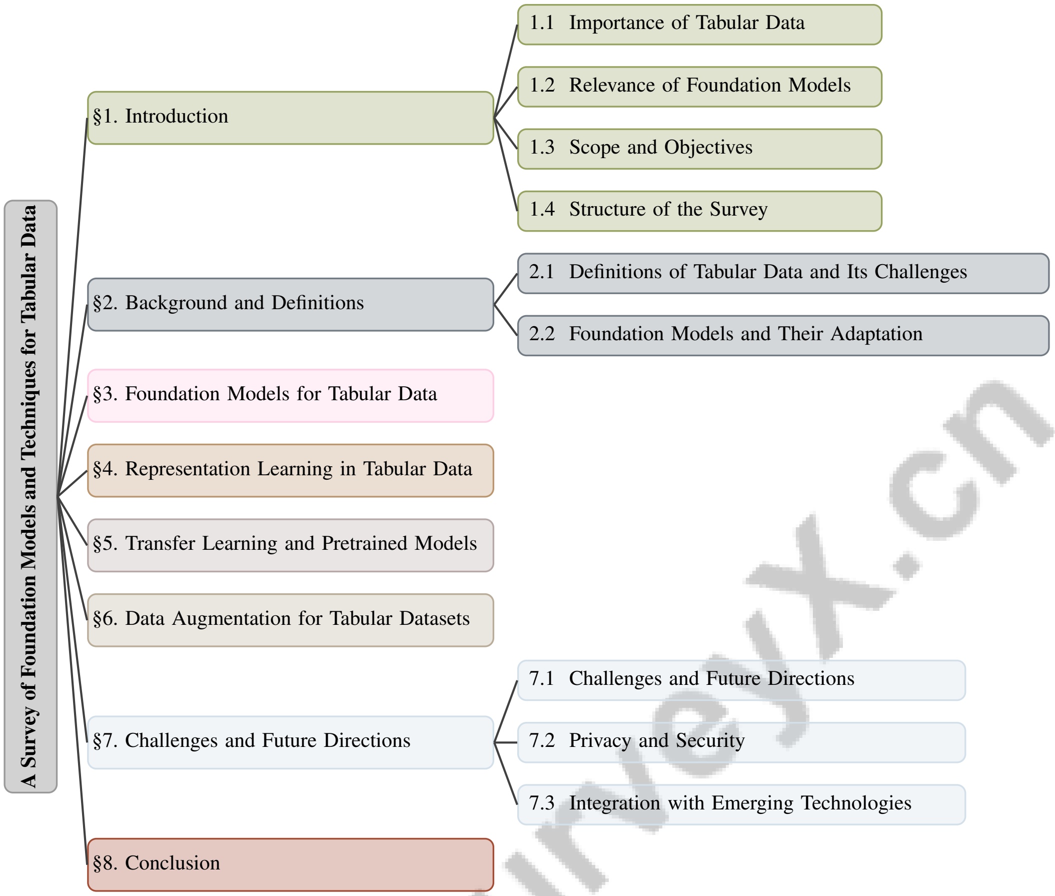

A Survey of Foundation Models and Techniques for Tabular Data
§1. Introduction
1.1 Importance of Tabular Data
1.2 Relevance of Foundation Models
1.3 Scope and Objectives
1.4 Structure of the Survey
§2. Background and Definitions
2.1 Definitions of Tabular Data and Its Challenges
2.2 Foundation Models and Their Adaptation
§3. Foundation Models for Tabular Data
§4. Representation Learning in Tabular Data
§5. Transfer Learning and Pretrained Models
§6. Data Augmentation for Tabular Datasets
§7. Challenges and Future Directions
7.1 Challenges and Future Directions
7.2 Privacy and Security
7.3 Integration with Emerging Technologies
§8. Conclusion

Figure 1: chapter structure

The significance of tabular data in advancing machine learning applications drives innovation and necessitates ongoing exploration to fully leverage its potential.

1.2 Relevance of Foundation Models

Foundation models are pivotal in advancing machine learning, particularly concerning tabular data. Pre-trained on extensive datasets, these models enhance the efficiency and performance of Multi-Task Learning (MTL) across various domains [10]. Their ability to automate complex tasks and generate human-interpretable concepts bridges causal representation learning and foundation models [9], which is particularly crucial in healthcare, where understanding model predictions can significantly influence clinical decisions [11].

Foundation models tackle persistent challenges such as data sparsity and mixed feature types in tabular datasets [5]. Utilizing methods like the Performance in a Graph (PinG) approach, these models employ Graph Neural Networks (GNN) to capture intricate relationships between features and samples, thereby overcoming limitations of existing techniques [4]. Tools like whyqd, a schema-focused toolkit, facilitate the integration of foundation models into tabular data tasks, enhancing data interoperability [8].

Beyond technical capabilities, foundation models prioritize data privacy and security, making them suitable for sensitive applications [12]. Large Language Models (LLMs), a subset of foundation models, can automate data analysis for tabular data, democratizing access to advanced analytical tools for nonspecialists [7].

Foundation models significantly enhance the processing, interpretation, and visualization of tabular data. Their ability to model complex relationships, improve diagnostic precision, and promote user engagement establishes them as essential for advancing machine learning applications across various fields, including digital pathology and interpretable machine learning. Emerging techniques leveraging human feedback to define transparent concepts in high-dimensional data further facilitate intuitive interactions with machine learning models, ensuring clarity and predictive accuracy [13, 9, 14, 15, 16].

1.3 Scope and Objectives

This survey aims to comprehensively examine foundation models and techniques tailored for tabular data, focusing on their construction and application across various domains. By emphasizing deep learning architectures, the survey seeks to bridge knowledge gaps in their application, particularly in industries where tabular data is critical [17]. It also explores adaptation strategies for domain-specific foundation models, highlighting their architecture and practical applications across different sectors [18].

In healthcare, the survey addresses the application of foundation models in areas such as natural language processing, medical image analysis, and graph learning, underscoring their transformative potential in clinical settings [19]. It also tackles limitations of existing pretrained language models in capturing data visualization semantics, proposing novel approaches for converting natural language queries into visual representations [20].

A key objective is to analyze the capabilities and limitations of clinical foundation models, particularly concerning Electronic Medical Records (EMRs), to enhance their effectiveness in healthcare applications [11]. Additionally, the survey proposes a new framework, Generative Tabular Learning (GTL), to address the inadequate generalization capabilities of existing tabular deep learning methods [1].

Lastly, the survey investigates the use of Large Language Models (LLMs) for tabular data analysis, focusing on the JarviX platform to democratize access to advanced analytical tools and empower nonspecialists in data-driven decision-making [7]. Through these objectives, the survey seeks to advance the understanding and application of foundation models in tabular data analysis, fostering innovation and enhancing model performance across various domains.

1.4 Structure of the Survey

The survey is systematically organized into key sections, each addressing critical aspects of foundation models and their application to tabular data. The introduction outlines the significance of tabular data across industries and the relevance of foundation models in addressing unique challenges. A comprehensive background section defines essential concepts such as tabular data, foundation models, and representation learning, providing context for subsequent discussions.

The core sections delve into the deployment of foundation models tailored for tabular data, examining their performance, scalability, and integration with pretrained models to enhance task-specific outcomes. A dedicated section explores representation learning techniques that automatically discover meaningful data representations, emphasizing recent advancements and their implications for model interpretability.

The role of transfer learning and pretrained models in enhancing machine learning tasks on tabular datasets is critically evaluated, focusing on successful applications and challenges associated with cross-table pretraining. Recent research highlights innovative frameworks like CM2, which employs a semantic-aware tabular neural network and a novel pretraining objective called prompt Masked Table Modeling (pMTM) to effectively address complexities in heterogeneous tabular data. The TabRet model further exemplifies the effectiveness of retokenizing feature embeddings for unseen columns, achieving top performance in healthcare classification tasks. These developments underscore the need for further exploration, particularly in overcoming fixed-schema limitations and advancing tabular data pretraining [21, 22]. The survey also reviews various data augmentation techniques to improve model accuracy and generalization through synthetic data generation.

In the latter sections, current challenges and future research directions are identified, including privacy and security concerns and the integration of emerging technologies with foundation models. The conclusion synthesizes key findings, reflecting on the potential of these models to advance tabular data analysis and the necessity for ongoing research and innovation in this domain.The following sections are organized as shown in Figure 1.

2 Background and Definitions

2.1 Definitions of Tabular Data and Its Challenges

Tabular data is structured in rows and columns, with each row representing an instance and each column a feature, prevalent in sectors like healthcare, finance, and retail. Its machine learning application is hindered by challenges such as non-standardized data integration from diverse sources, complicating model deployment [8]. The manual creation of interactive features further impedes robust model development [5]. High dimensionality increases overfitting risks, especially with limited training instances [1], and imbalanced datasets complicate effective training, as conventional loss functions like MSE are inadequate for skewed categorical variables [6]. Creating interpretable representations to elucidate data impact on model predictions is crucial, particularly in critical fields like healthcare [9].

In High Performance Computing (HPC), execution time prediction is challenged by missing data and the need to model cross-sample relationships [4]. Advanced data analytics sophistication presents further challenges for nonspecialists, complicating the utilization of Large Language Models (LLMs) for tabular data analysis [7]. Addressing these challenges necessitates innovative machine learning frameworks capable of managing tabular data complexities, enhancing model interpretability, and delivering accurate, generalizable predictions across diverse applications.

2.2 Foundation Models and Their Adaptation

Adapting foundation models to tabular data involves innovative methodologies to address the unique challenges of structured datasets. The Dynamic Weighted Tabular Method (DWTM) exemplifies this by dynamically assigning feature weights based on statistical significance, enhancing CNN classification accuracy [23]. Many approaches fail to capture intra-field information and non-linear interactions, limiting classification task effectiveness [24]. The PTab framework addresses this by converting tabular data into text, allowing pre-trained language models to derive contextual representations [25].

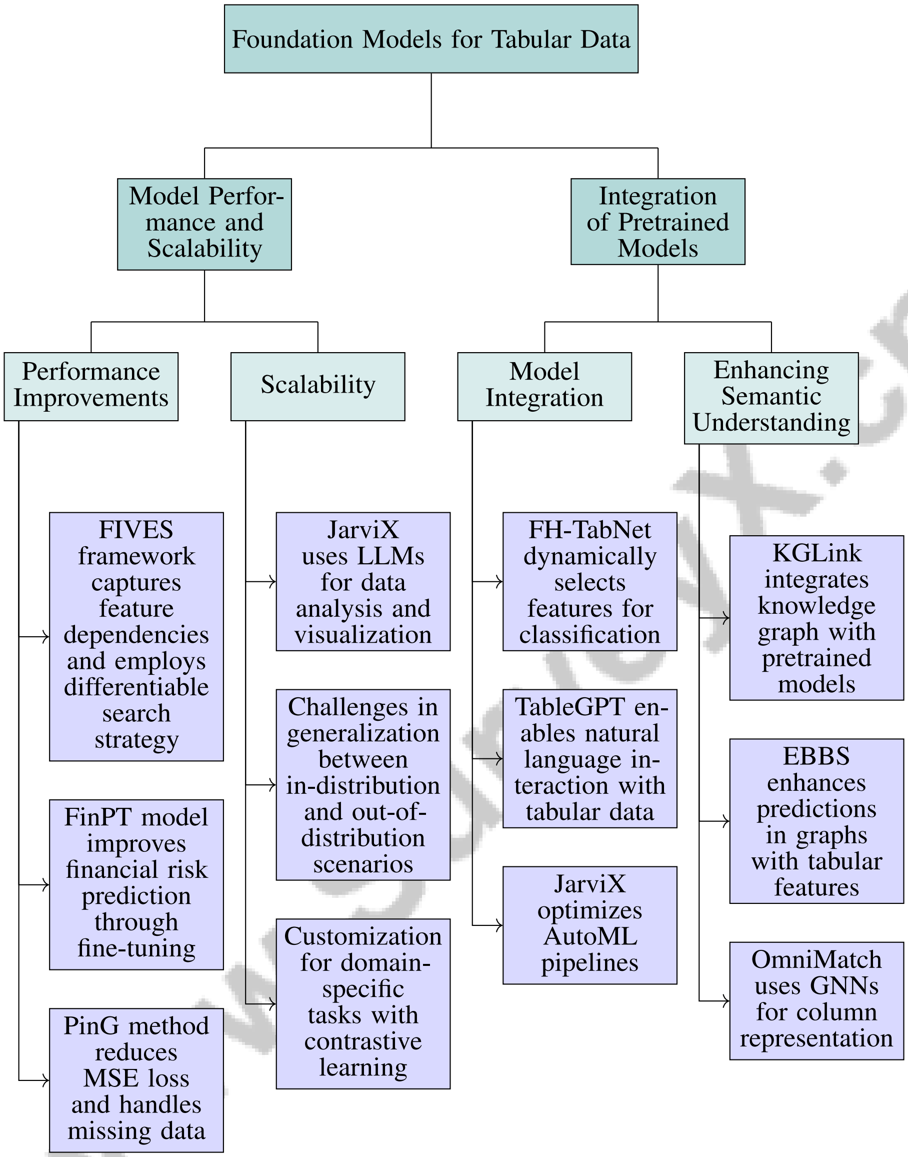
LLM integration into tabular deep learning is hindered by domain-specific knowledge comprehension challenges [1]. Multi-Task Learning (MTL) techniques mitigate these by emphasizing regularization, relationship learning, feature propagation, optimization, and pre-training [10]. These techniques optimize foundation model performance across multiple tasks, enhancing applicability to tabular data. The Whyqd toolkit enhances data interoperability through schema-to-schema mappings [8], while the FIVES framework models feature interactions as edges in a feature graph, optimizing graph structure and prediction models to capture complex tabular data relationships [5].

The Scaled Autoencoder for Mixed Tabular Datasets (SAM) incorporates a balanced loss function to address imbalanced datasets, ensuring robust performance [6]. Organizing methods by data types and representation stages, like discretization and binarization, supports foundation model adaptation to tabular data [9]. Adapting foundation models for tabular data requires dynamic feature weighting, contextual representation learning, schema-focused mapping, and innovative feature interaction modeling. These initiatives are crucial for overcoming existing methodology limitations and significantly enhancing foundation model adaptability across diverse tabular datasets. By leveraging advanced capabilities in natural language processing, content generation, and multimodal integration, these efforts improve performance in tasks such as table-class detection, column-type annotation, and join-column prediction, surpassing traditional task-specific models. The development of specialized models like LaTable addresses the complexities of heterogeneous feature spaces in tabular data, enabling more effective applications across various domains [18, 26, 27, 28].

In recent years, the development of foundation models has significantly transformed the landscape of machine learning, particularly in the domain of tabular data. These models have not only improved performance but have also enhanced scalability and the integration of pretrained models. Figure 2 illustrates the hierarchical structure of these foundation models, categorizing key advancements in model performance, scalability, and the role of pretrained models. Specifically, the model performance section highlights frameworks such as FIVES and PinG, which are pivotal in enhancing both performance and scalability. Furthermore, the integration section underscores the importance of pretrained models in improving semantic understanding and facilitating more intuitive data inter-

AI-generated, for reference only.

actions. This comprehensive framework serves as a foundation for understanding the multifaceted advancements in this rapidly evolving field.



Foundation Models for Tabular Data
Model Performance and Scalability
Integration of Pretrained Models
Performance Improvements
Scalability
Model Integration
Enhancing Semantic Understanding
FIVES framework captures feature dependencies and employs differentiable search strategy
FinPT model improves financial risk prediction through fine-tuning
PinG method reduces MSE loss and handles missing data
JarviX uses LLMs for data analysis and visualization
Challenges in generalization between in-distribution and out-of-distribution scenarios
Customization for domain-specific tasks with contrastive learning
FH-TabNet dynamically selects features for classification
TableGPT enables natural language interaction with tabular data
JarviX optimizes AutoML pipelines
KGLink integrates knowledge graph with pretrained models
EBBS enhances predictions in graphs with tabular features
OmniMatch uses GNNs for column representation

Figure 2: This figure illustrates the hierarchical structure of foundation models for tabular data, categorizing key advancements in model performance, scalability, and the integration of pretrained models. The model performance section highlights frameworks like FIVES and PinG that enhance performance and scalability. The integration section emphasizes the role of pretrained models in improving semantic understanding and facilitating intuitive data interactions.

3 Foundation Models for Tabular Data

3.1 Model Performance and Scalability

Foundation models have markedly improved the performance and scalability of tabular data applications, often surpassing traditional machine learning methodologies. The FIVES framework exemplifies these advancements by capturing feature dependencies and employing a differentiable search strategy to identify useful interactions, thereby enhancing performance in complex datasets [5]. Similarly, the FinPT model demonstrates consistent performance improvements over baseline

AI-generated, for reference only.

Benchmark	Size	Domain	Task Format	Metric
FinBench[3]	333,000	Financial Risk Prediction	Binary Classification	F1-score
TableBench[29]	886	Numerical Reasoning	Table Question Answering	ROUGE-L
FoundTS[30]	1,000,000	Electricity	Forecasting	MAE, MSE
RetFound[31]	1,416	Ophthalmology	Binary Classification	AUC
CTGAN[32]	15	Synthetic Data Generation	Data Synthesis	Likelihood Fitness, Machine Learning Efficacy
FL-GAN[33]	15,000	Finance	Classification	JSD, WD
REIN[34]	1,000,000	Business	Classification	F1-score, IoU
MIA-TDS[35]	100,000	Tabular Data Synthesis	Membership Inference	AUROC

Table 1: The table provides a comprehensive overview of various benchmarks utilized in foundation model research, detailing their size, domain, task format, and evaluation metrics. This compilation highlights the diversity in applications and the specific metrics used to assess model performance across different domains.

models, particularly in fine-tuning large foundation models for financial risk prediction, showcasing their adaptability to domain-specific tasks [3].

The Performance in a Graph (PinG) method underscores the efficacy of foundation models by achieving significant reductions in Mean Square Error (MSE) loss while adeptly handling missing data, thus boosting regression task efficiency [4]. This capability is crucial for the scalability of foundation models, enabling efficient processing of large-scale, intricate tabular datasets. Table 1 presents a detailed comparison of representative benchmarks used in evaluating the performance and scalability of foundation models across various domains.

JarviX, a platform leveraging Large Language Models (LLMs), further illustrates the scalability of foundation models through precise data analysis and visualization, offering user-friendly automation and tailored insights for extensive tabular data management [7]. This scalability is vital across various sectors where rapid and accurate data processing is essential.

Despite these advancements, challenges persist in enhancing generalization capabilities, particularly in bridging the performance gap between in-distribution and out-of-distribution scenarios. The limitations of general-purpose models often restrict their ability to capture domain-specific patterns and requirements. As research focuses on customizing foundation models for specific domains, understanding contrastive learning and its application to heterogeneous data becomes essential for enhancing model robustness across diverse tasks [18, 36, 28]. Ongoing research is crucial to addressing these challenges and fully exploiting the potential of foundation models across varied applications.

3.2 Integration of Pretrained Models

The integration of pretrained models into tabular data tasks has significantly improved the performance and versatility of machine learning applications. The FH-TabNet model exemplifies this integration by dynamically selecting features for each sample during classification, thereby enhancing the model's ability to differentiate categories and improve overall performance [37].

As illustrated in Figure 3, the integration of pretrained models into tabular data tasks encompasses various applications, enhancements in semantic understanding, and contributions to visualization and interaction. This figure categorizes key methods and innovations, showcasing their impact on improving data analysis and interaction capabilities. For instance, TableGPT highlights the integration of pretrained models by enabling users to query, manipulate, and visualize tabular data via a natural language interface [38]. This approach leverages pretrained models to facilitate intuitive data interactions, broadening access to advanced data analysis. Furthermore, the integration of LLMs into automated machine learning (AutoML) pipelines, as demonstrated by the JarviX platform, optimizes data analysis processes and democratizes access to sophisticated analytical tools [7].

Enhancing semantic understanding through pretrained models is also critical. KGLink integrates knowledge graph information with pretrained deep learning models to improve the semantic annotation of table columns, enriching contextual understanding [39]. Additionally, methods like EBBS combine graph-aware propagation with boosting to enhance predictions in graphs with tabular features [40].

AI-generated, for reference only.

Graph Neural Networks (GNNs) play a pivotal role in this integration, with the OmniMatch method using GNNs to learn column representations based on pairwise similarities, thereby improving data integration processes by identifying joinable columns [41]. This underscores the significance of pretrained models in enhancing tabular data interoperability.

The integration also extends to visualization and interaction, where models facilitate chart generation, visual enhancements, and user interactions, thereby improving exploratory capabilities [28]. Additionally, methods like Oases optimize training schedules of foundation models by maximizing communication and computation overlap, enhancing training efficiency [42].

Incorporating pretrained models into tabular tasks enhances performance through advanced techniques, such as generative AI for data augmentation and innovative architectures like TabNet and CM2. These methods improve data understanding via sophisticated feature representations and attention mechanisms while democratizing access to analytical tools, ultimately advancing the field of tabular data analysis and addressing challenges related to data scarcity and heterogeneous structures [21, 16, 17, 22].

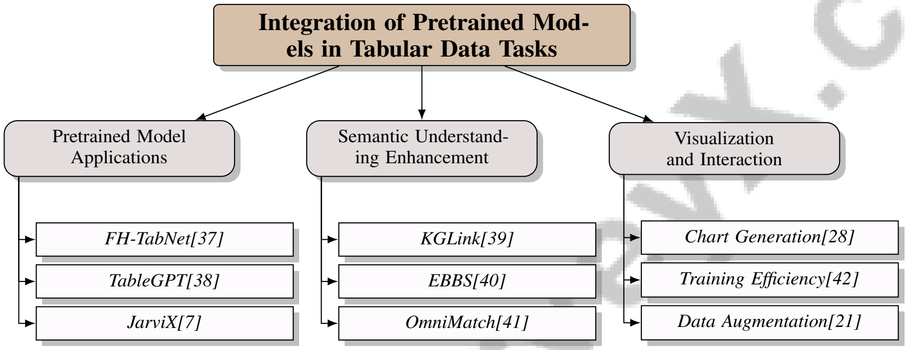

Integration of Pretrained Models in Tabular Data Tasks
Pretrained Model Applications
Semantic Understanding Enhancement
Visualization and Interaction
FH-TabNet[37]
TableGPT[38]
JarviX[7]
KGLink[39]
EBBS[40]
OmniMatch[41]
Chart Generation[28]
Training Efficiency[42]
Data Augmentation[21]

Figure 3: This figure illustrates the integration of pretrained models into tabular data tasks, highlighting their applications, enhancements in semantic understanding, and contributions to visualization and interaction. It categorizes key methods and innovations, showcasing their impact on improving data analysis and interaction capabilities.

4 Representation Learning in Tabular Data

4.1 Representation Learning and Interpretability

Representation learning is crucial for enhancing the interpretability of machine learning models applied to tabular data. Methods like Adaptive Explainable Visualization (ADAVIS) and TabAttention utilize attention mechanisms to provide transparent visualization and highlight relevant features, improving model interpretability across both imaging and tabular contexts [43, 44]. In high-stakes fields such as healthcare, foundation models (FMs) have demonstrated improved accuracy and efficiency, underscoring the importance of interpretability [19]. The Multimodal Contrastive Learning approach, incorporating a tabular attention mechanism, ranks feature importance and aids in understanding complex disease scenarios like Alzheimer's [45].

Concept-based models enhance interpretability by allowing direct labeling of concept features, increasing efficiency in representation learning [13]. The LEGATO method supports end-to-end learning of information aggregation strategies, improving interpretability by leveraging actual data relationships [46]. Techniques like Local Interpretable Model-agnostic Explanations (LIME) and Denoising Autoencoder provide reliable explanations and expected values for erroneous cells, respectively, aiding domain experts in understanding model behavior and correcting errors in mixed-type data [47, 48].

Advanced methods such as DAVID capture nonlinear relationships, essential for representation learning in tabular datasets [49]. The TabPFN method transforms time series data into tabular format, capturing temporal relationships and predicting future target values, enhancing interpretability in time-dependent data [50]. The Network of Networks (NON) approach effectively models complex interactions non-linearly, enhancing classification accuracy [24]. These techniques collectively emphasize the significance of interpretability in representation learning, fostering a deeper understanding of model predictions and promoting trust in machine learning applications.

4.2 Advancements and Implications

Recent advancements in representation learning for tabular data have significantly improved the ability to capture complex relationships and enhance model performance. The TabPFN method exemplifies this progress by transforming time series data into a tabular format, enabling the capture of temporal relationships and future target value predictions [50]. The integration of attention mechanisms, as seen in the TabAttention framework, emphasizes relevant features through multi-head self-attention, enhancing interpretability and effectiveness in representation learning [44].

Figure 4 illustrates the key advancements and implications in representation learning for tabular data, healthcare applications, and interpretable machine learning models, highlighting methods like TabPFN, TabAttention, and multimodal contrastive learning, as well as their impact on healthcare and interpretability. In healthcare, foundation models (FMs) have demonstrated improved accuracy and efficiency, reinforcing the importance of interpretability in high-stakes environments [19]. The Multimodal Contrastive Learning approach, incorporating a tabular attention mechanism, ranks feature importance in complex disease contexts, such as Alzheimer's, providing insights into disease mechanisms and enhancing diagnostic accuracy [45].

Concept-based models have improved interpretability by enabling direct labeling of concept features, increasing efficiency in representation learning [13]. The LEGATO method supports end-to-end learning of information aggregation strategies, enhancing interpretability [46].

Advancements in interpretable machine learning models, particularly generalized additive models (GAMs), hold significant implications across various domains. These models enhance both accuracy and interpretability, challenging the notion that high performance is exclusive to black-box models, while fostering transparency and trust. As organizations increasingly rely on data-driven decision-making, the necessity for models that are both effective and understandable becomes paramount. Robust evaluation methods for explainable AI (XAI) ensure the fidelity of interpretations aligns closely with the underlying predictive mechanisms, enhancing confidence in their deployment across diverse fields [14, 51]. By improving the capacity to capture and interpret complex data relationships, these techniques pave the way for informed decision-making processes, especially in areas where model predictions carry significant consequences. As representation learning continues to evolve, its influence on the analysis and interpretation of tabular data is expected to expand, driving innovation and broadening the applicability of machine learning solutions.

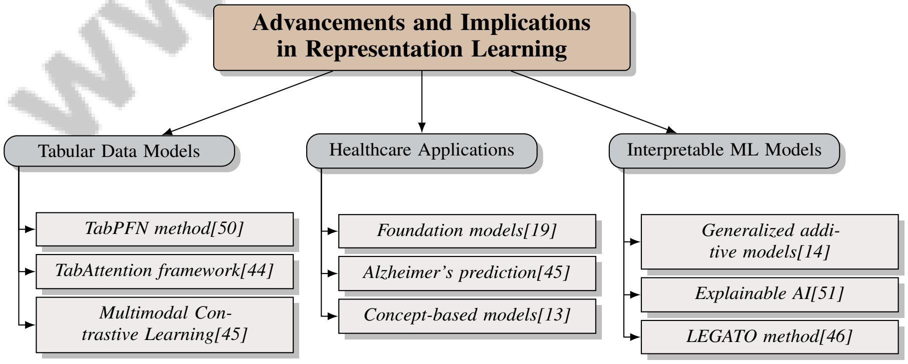

Advancements and Implications
in Representation Learning
Tabular Data Models
Healthcare Applications
Interpretable ML Models
TabPFN method[50]
TabAttention framework[44]
Multimodal Contrastive Learning[45]
Foundation models[19]
Alzheimer's prediction[45]
Concept-based models[13]
Generalized additive models[14]
Explainable AI[51]
LEGATO method[46]

Figure 4: This figure illustrates the key advancements and implications in representation learning for tabular data, healthcare applications, and interpretable machine learning models, highlighting methods like TabPFN, TabAttention, and multimodal contrastive learning, as well as their impact on healthcare and interpretability.

5 Transfer Learning and Pretrained Models

5.1 Pretrained Models and Transfer Learning

Pretrained models are pivotal in transfer learning, enhancing machine learning for tabular data by utilizing knowledge from large datasets, thus reducing computational demands and improving performance. Benchmarks of models like GPT-2, T5, and LLaMA in financial risk prediction demonstrate the effectiveness of pretrained models in boosting predictive accuracy [3]. Their adaptability supports fine-tuning for domain-specific tasks.

The JarviX platform exemplifies the application of pretrained models in tabular data, using Large Language Models (LLMs) to optimize data processing and interpretation [7]. By automating analysis, JarviX democratizes advanced data tools, aiding nonspecialists in data-driven decision-making across sectors.

In data-scarce scenarios, integrating pretrained models into transfer learning frameworks is advantageous. The "cliff-learning" concept highlights significant performance gains with minor data increases, crucial in high-stakes contexts where interpretability and performance are vital [52, 14].

The TabRet model, using a retokenizing strategy within a Transformer-based framework, adeptly manages unseen columns in tabular datasets by updating tokenizers, ensuring adaptability and sustained performance [22]. Advanced representation learning techniques that leverage sample-feature relationships enhance robustness in tabular predictions [4].

As illustrated in Figure 5, the hierarchical structure of pretrained models and transfer learning encompasses model benchmarks, platforms and methods, and evaluation frameworks within the context of machine learning for tabular data. This visual representation underscores the interconnectedness of these elements, providing a comprehensive overview of how pretrained models facilitate effective learning.

Benchmarks like DataDEL, offering novel datasets and evaluation metrics, advance pretrained models in transfer learning by enabling comprehensive assessments of data-efficient learning and emphasizing the balance between interpretability and performance [53]. These frameworks are crucial in applications with significant predictive implications, fostering innovation and enhancing decision-making across diverse fields.

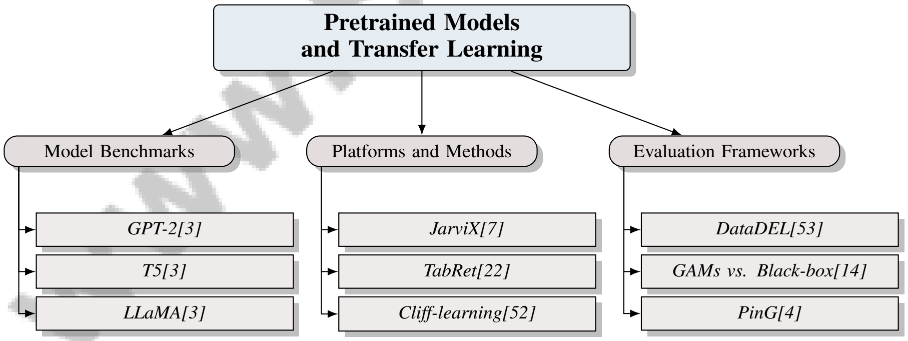


Figure 5: This figure illustrates the hierarchical structure of pretrained models and transfer learning, highlighting model benchmarks, platforms and methods, and evaluation frameworks in the context of machine learning for tabular data.

5.2 Challenges in Cross-Table Pretraining

Cross-table pretraining encounters challenges that can limit its effectiveness. A key limitation is the small parameter size of models like CM2, which may not handle complex tasks as effectively as leading language models [21]. This necessitates exploring strategies to enhance model capacity and scalability, potentially through architectural innovations or efficient parameter use.

The dependency on pretrained model quality and transformation selection during pretraining is another hurdle. The method's effectiveness can be compromised if initial models lack robustness or if transformations do not align with target tasks [54]. This underscores the need for robust evaluation frameworks and criteria for selecting pretrained models and transformations suited to specific tabular data characteristics.

The cliff-learning phenomenon, indicating substantial performance gains with slight data increases, may not apply universally across all scenarios [52]. This variability complicates predictions about cross-table pretraining effectiveness, especially in data-scarce domains. Addressing this requires further research to understand cliff-learning conditions and develop adaptive strategies generalizing across scenarios.

6 Data Augmentation for Tabular Datasets

6.1 Data Augmentation Techniques

Data augmentation is vital for improving machine learning models' performance and generalization on tabular datasets, particularly by addressing overfitting and data scarcity. Techniques such as generating synthetic samples by mapping real data neighborhoods into their convex space allow for classification against other samples while preserving the dataset's intrinsic structure [55]. The evolution of data synthesis is marked by four technological eras: Expert Knowledge, Direct Training, Pre-train then Fine-tune, and Foundation Models without Fine-tuning, each reflecting advances in computational methods and model architectures [56].

Generative models like the Tabular Generative Adversarial Network (TGAN) represent innovative data augmentation techniques, generating high-quality synthetic tables by modeling relationships among mixed variable types and preserving correlations [57]. This capability is crucial for maintaining the statistical properties of the original dataset, ensuring synthetic data remains realistic and useful for training. Additionally, adaptive robust watermarking techniques ensure the integrity of the original dataset while embedding minimal modifications, essential for privacy-sensitive applications [58].

As illustrated in Figure 6, the landscape of data augmentation for tabular datasets is rapidly evolving, incorporating advanced methodologies such as generative AI to enhance original data through synthetic example generation or external data retrieval. This figure highlights the hierarchical categorization of data augmentation techniques, emphasizing generative models, watermarking techniques, and deep learning models as key approaches in enhancing tabular datasets. This evolution includes processes like error handling and schema matching, alongside augmentation methods categorized into retrieval-based and generation-based approaches, addressing different granularity levels. Recent advancements in deep learning models like TabNet and SAINT utilize attention mechanisms and hybrid architectures tailored for tabular data challenges, further improving model performance and generalization [17]. Models such as TabRet and TGAN illustrate the potential of pre-training and generative adversarial networks to synthesize high-quality tabular data, addressing data scarcity and enhancing robustness across various applications, including healthcare and finance [57, 16, 22].

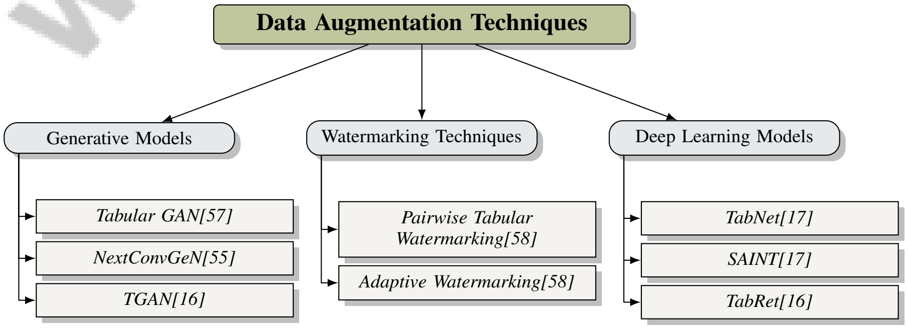


Figure 6: This figure illustrates the hierarchical categorization of data augmentation techniques, highlighting generative models, watermarking techniques, and deep learning models as key approaches in enhancing tabular datasets.

6.2 Generative and Synthetic Data Techniques

Generative models and synthetic data techniques are crucial for augmenting tabular datasets, effectively addressing data scarcity and enhancing model performance. Methods like NextConvGeN generate synthetic samples within the convex space of data neighborhoods and use a discriminator to classify these samples against randomly sampled batches, maintaining structural integrity and diversity [55]. The evolution of data synthesis methodologies, structured into stages beginning with Synthetic Data Generation and followed by Post-processing, emphasizes quality, label consistency, and adherence to the original data distribution [56].

The TGAN model exemplifies innovation by integrating Long Short-Term Memory (LSTM) networks with attention mechanisms, enabling sequential data generation and effective handling of multimodal distributions prevalent in tabular datasets [57]. Attention mechanisms enhance TGAN's ability to capture complex dependencies, improving the quality and applicability of synthetic data.

Generative and synthetic data techniques enhance machine learning models by providing high-quality training data, addressing challenges like data scarcity, privacy concerns, and the need for improved diversity and balance in datasets. By leveraging advanced methods like deep generative models and tabular data augmentation, these techniques produce data closely mirroring real-world distributions, improving model performance across various applications while adhering to privacy standards like differential privacy. The strategic application of synthetic data can significantly enhance analytical accuracy and robustness, as demonstrated by frameworks integrating statistical methods with synthetic data generation, redefining the data science landscape [59, 56, 16, 60]. These techniques mitigate limitations posed by small or imbalanced datasets, paving the way for more robust and generalizable machine learning solutions across diverse domains.

7 Challenges and Future Directions

7.1 Challenges and Future Directions

Integrating foundation models with tabular data poses significant challenges that require innovative research to improve their performance and applicability. A major issue is the high computational demand of large language models (LLMs), leading to greater operational costs compared to traditional machine learning methods. Additionally, LLMs are constrained by context length, which can limit their effectiveness in specific applications. Despite these challenges, LLMs have shown potential as weak learners in boosting algorithms, enhancing classification tasks with limited data. Foundation models excel in data discovery and exploration tasks, often surpassing traditional models and human experts, indicating their broader potential in data management [61, 26]. Addressing these issues requires advancements in model compression and the development of efficient architectures to utilize LLMs while reducing computational burdens.

Figure 7 illustrates the key challenges and future directions in integrating foundation models with tabular data. It highlights three main areas: improving model efficiency through techniques like model compression and LLM boosting, ensuring data privacy and security with federated learning and privacy-preserving techniques, and enhancing data integration by focusing on schema transformation, multimodal fusion, and feature interaction.

In federated learning, balancing privacy and data utility remains a critical concern. Future research should focus on improving privacy-preserving techniques and model compression in Federated Foundation Models, ensuring data security while maintaining performance [12]. Integrating hypernetworks with other architectures could enhance data fusion capabilities, particularly in multimodal applications like reverse conditioning of Electronic Health Records (EHR) analysis on imaging data [2].

Transforming diverse data formats into a unified schema is complex, especially in schema-based contexts. Future research should aim to simplify these transformation processes and reduce the associated learning curve to facilitate broader adoption [8]. Developing flexible frameworks in Multi-Task Learning (MTL) is crucial for enhancing pretrained models' capabilities, enabling efficient handling of multiple tasks [10].

Generalization capabilities are vital, especially when applying foundation models to medium-sized tabular datasets prone to overfitting. Future research should explore additional inductive biases, such as those found in tree-based models, and extend benchmarks to accommodate various dataset sizes, thereby improving model robustness and adaptability [5]. Enhancing feature generation completeness and expanding methods like FIVES to other data types could further improve predictive accuracy [5].

In financial risk prediction, future research could explore additional financial risks or enhance dataset quality and balancing methods to address current challenges [3]. Optimizing graph construction processes and exploring additional graph-based techniques are promising directions for improving predictive accuracy across applications [4].

Developing robust and trustworthy interpretable representations that adapt to diverse data types is essential for advancing explainable AI [9]. Enhancing LLM personalization and expanding compatibility with various data types are also critical future directions to improve analytical capabilities [7]. Focusing on these research directions will enhance the robustness, interpretability, and applicability of foundation models in tabular data analysis, facilitating informed decision-making and innovation in machine learning applications across domains.

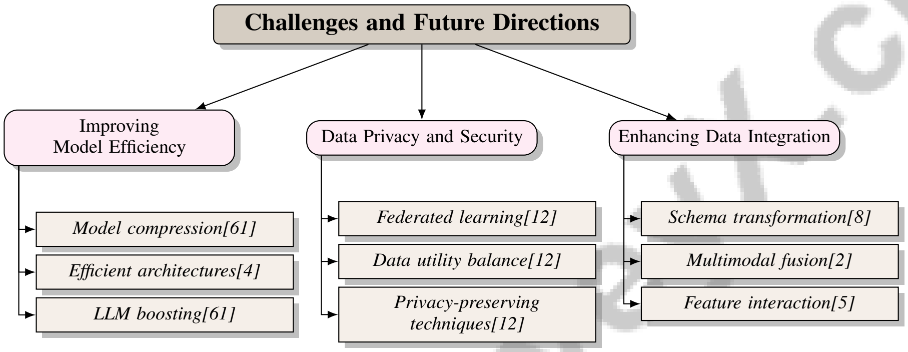

Challenges and Future Directions
Improving Model Efficiency
Data Privacy and Security
Enhancing Data Integration
Model compression[61]
Efficient architectures[4]
LLM boosting[61]
Federated learning[12]
Data utility balance[12]
Privacy-preserving techniques[12]
Schema transformation[8]
Multimodal fusion[2]
Feature interaction[5]

Figure 7: This figure illustrates the key challenges and future directions in integrating foundation models with tabular data. It highlights three main areas: improving model efficiency through techniques like model compression and LLM boosting, ensuring data privacy and security with federated learning and privacy-preserving techniques, and enhancing data integration by focusing on schema transformation, multimodal fusion, and feature interaction.

7.2 Privacy and Security

Privacy and security issues are paramount in handling tabular data, especially in sectors like healthcare where sensitive information is prevalent. The General Data Protection Regulation (GDPR) underscores the importance of interpretability in machine learning models to ensure data privacy and security, highlighting the need for transparent AI systems [47]. The deployment of models like TableGPT in private environments underscores the critical need for maintaining data privacy while leveraging advanced analytical capabilities [38].

Federated Foundation Models (FFMs) offer a promising solution by enabling data privacy and communication efficiency, minimizing risks associated with data sharing [12]. These models support collaborative learning across multiple data sources without centralizing sensitive data, thus preserving privacy while enhancing performance.

Integrating differential privacy algorithms is another approach to safeguarding privacy in machine learning applications. The choice of algorithm significantly impacts model performance and privacy protection, necessitating careful consideration to balance these factors [35]. Moreover, the requirement for FDA-like disclosures on data and algorithms used in Foundation Models (FMs) is crucial for ensuring transparency and accountability in medical AI systems [62].

Enhancing the interpretability and robustness of models trained on heterogeneous data is essential for creating comprehensive benchmarks and enhancing the trustworthiness of AI systems [36]. The interplay between data mining practices and Explainable Artificial Intelligence (XAI) is vital for establishing a transparent AI landscape that considers privacy and security [63].

Finally, understanding and mitigating bias in Foundation Models is critical for ensuring equitable outcomes in healthcare and maintaining trust in medical AI systems [31]. These efforts are complemented by advancements in graph construction and the development of extensive datasets, necessary for validating findings and enhancing privacy and security in machine learning applications [64]. Addressing these challenges will advance secure and privacy-conscious AI solutions, fostering innovation and trust in the use of tabular data across various domains.

7.3 Integration with Emerging Technologies

The integration of emerging technologies with foundation models for tabular data is crucial for enhancing machine learning capabilities across diverse domains. Hybrid architectures that merge neural networks with traditional machine learning models are particularly promising, leveraging the strengths of both approaches to address challenges posed by tabular data and improve model performance and scalability [17]. In healthcare, incorporating clinical foundation models (FMs) into existing workflows has the potential to enhance diagnostic accuracy and patient outcomes, with future research focusing on including additional data sources to broaden the applicability of models like FH-TabNet in various healthcare settings [37].

The integration of Tabular Data Augmentation (TDA) techniques with emerging technologies such as large language models (LLMs) and hybrid approaches that combine retrieval and generation methods is another promising direction, enhancing the robustness and generalization capabilities of machine learning models, particularly in data-scarce scenarios [16]. Additionally, exploring improvements in self-supervised learning within frameworks like PTab and addressing representation challenges of numerical values in textual modalities are crucial areas for future research [25].

Future research could also focus on integrating Generative Tabular Learning (GTL) with traditional models to leverage their complementary strengths, facilitating enhanced performance and interpretability [1]. The integration of linked data into schemas, as suggested by developing auditable and reusable crosswalks, aligns with emerging technologies and enhances foundation models' capabilities for tabular data, providing a robust framework for data interoperability and analysis [8].

Further optimizations in constraint extraction processes and the development of additional operators to enhance frameworks like Morph-CSV are also critical for improving the efficiency and effectiveness of data integration and analysis [65]. By focusing on these research directions, the integration of emerging technologies with foundation models can drive innovation and expand the applicability of machine learning solutions in analyzing tabular data, ultimately fostering informed decision-making and advancing the field.

8 Conclusion

This survey underscores the profound impact of foundation models on tabular data analysis, showcasing their ability to drive significant advancements across diverse sectors, notably in healthcare. The FH-TabNet framework exemplifies this advancement by achieving remarkable F1-scores in the classification of rare familial hypercholesterolemia subcategories, thereby enhancing diagnostic precision and operational workflows. The integration of generative and synthetic data methodologies, as illustrated by the Syn framework, effectively mitigates issues of data scarcity and privacy, thereby refining the accuracy of statistical analyses applied to synthetic datasets. Such methodologies are crucial in extending the reach of machine learning solutions in data-constrained environments. Furthermore, the HoneyBee framework highlights the efficacy of combining multimodal data to generate superior embeddings, significantly boosting machine learning outcomes in oncology. This integration of diverse data types is pivotal for unlocking the full potential of foundation models, facilitating more holistic data analyses. Beyond traditional data analysis, the potential of foundation models extends to fields such as robot learning, where enhanced multimodal interactions can further optimize learning processes. This survey highlights the indispensable role of foundation models in advancing tabular data analysis, emphasizing the need for continuous research and innovation to address existing challenges and expand their applicability across various fields. The ongoing evolution and assimilation of these models are set to propel machine learning forward, supporting informed decision-making and improving the robustness and clarity of analytical frameworks.

References

[1] Xumeng Wen, Han Zhang, Shun Zheng, Wei Xu, and Jiang Bian. From supervised to generative: A novel paradigm for tabular deep learning with large language models, 2024.

[2] Daniel Duenias, Brennan Nichyporuk, Tal Arbel, and Tammy Riklin Raviv. Hyperfusion: A hypernetwork approach to multimodal integration of tabular and medical imaging data for predictive modeling, 2025.

[3] Yuwei Yin, Yazheng Yang, Jian Yang, and Qi Liu. Finpt: Financial risk prediction with profile tuning on pretrained foundation models, 2023.

[4] Tarek Ramadan, Ankur Lahiry, and Tanzima Z. Islam. Novel representation learning technique using graphs for performance analytics, 2024.

[5] Yuexiang Xie, Zhen Wang, Yaliang Li, Bolin Ding, Nezihe Merve Gürel, Ce Zhang, Minlie Huang, Wei Lin, and Jingren Zhou. Fives: Feature interaction via edge search for large-scale tabular data, 2021.

[6] Samuel Stocksieker, Denys Pommeret, and Arthur Charpentier. Boarding for iss: Imbalanced self-supervised: Discovery of a scaled autoencoder for mixed tabular datasets, 2024.

[7] Shang-Ching Liu, ShengKun Wang, Wenqi Lin, Chung-Wei Hsiung, Yi-Chen Hsieh, Yu-Ping Cheng, Sian-Hong Luo, Tsungyao Chang, and Jianwei Zhang. Jarvix: A llm no code platform for tabular data analysis and optimization, 2023.

[8] Gavin Chait. Auditable and reusable crosswalks for fast, scaled integration of scattered tabular data, 2024.

[9] Kacper Sokol and Peter Flach. Interpretable representations in explainable ai: From theory to practice, 2024.

[10] Jun Yu, Yutong Dai, Xiaokang Liu, Jin Huang, Yishan Shen, Ke Zhang, Rong Zhou, Eashan Adhikarla, Wenxuan Ye, Yixin Liu, Zhaoming Kong, Kai Zhang, Yilong Yin, Vinod Namboodiri, Brian D. Davison, Jason H. Moore, and Yong Chen. Unleashing the power of multi-task learning: A comprehensive survey spanning traditional, deep, and pretrained foundation model eras, 2024.

[11] Michael Wornow, Yizhe Xu, Rahul Thapa, Birju Patel, Ethan Steinberg, Scott Fleming, Michael A. Pfeffer, Jason Fries, and Nigam H. Shah. The shaky foundations of clinical foundation models: A survey of large language models and foundation models for emrs, 2023.

[12] Sixing Yu, J. Pablo Muñoz, and Ali Jannesari. Federated foundation models: Privacy-preserving and collaborative learning for large models, 2024.

[13] Isaac Lage and Finale Doshi-Velez. Learning interpretable concept-based models with human feedback, 2020.

[14] Sven Kruschel, Nico Hambauer, Sven Weinzierl, Sandra Zilker, Mathias Kraus, and Patrick Zschech. Challenging the performance-interpretability trade-off: An evaluation of interpretable machine learning models, 2024.

[15] Gustav Bredell, Marcel Fischer, Przemyslaw Szostak, Samaneh Abbasi-Sureshjani, and Alvaro Gomariz. The importance of downstream networks in digital pathology foundation models, 2024.

[16] Lingxi Cui, Huan Li, Ke Chen, Lidan Shou, and Gang Chen. Tabular data augmentation for machine learning: Progress and prospects of embracing generative ai, 2024.

[17] Shriyank Somvanshi, Subasish Das, Syed Aaqib Javed, Gian Antariksa, and Ahmed Hossain. A survey on deep tabular learning, 2024.

[18] Haolong Chen, Hanzhi Chen, Zijian Zhao, Kaifeng Han, Guangxu Zhu, Yichen Zhao, Ying Du, Wei Xu, and Qingjiang Shi. An overview of domain-specific foundation model: key technologies, applications and challenges, 2024.

[19] Wasif Khan, Seowung Leem, Kyle B. See, Joshua K. Wong, Shaoting Zhang, and Ruogu Fang. A comprehensive survey of foundation models in medicine, 2025.

[20] Zhuoyue Wan, Yuanfeng Song, Shuaimin Li, Chen Jason Zhang, and Raymond Chi-Wing Wong. Datavist5: A pre-trained language model for jointly understanding text and data visualization, 2024.

[21] Chao Ye, Guoshan Lu, Haobo Wang, Liyao Li, Sai Wu, Gang Chen, and Junbo Zhao. Towards cross-table masked pretraining for web data mining, 2024.

[22] Soma Onishi, Kenta Oono, and Kohei Hayashi. Tabret: Pre-training transformer-based tabular models for unseen columns, 2023.

[23] Md Ifraham Iqbal, Md. Saddam Hossain Mukta, and Ahmed Rafi Hasan. A dynamic weighted tabular method for convolutional neural networks, 2022.

[24] Yuanfei Luo, Hao Zhou, Weiwei Tu, Yuqiang Chen, Wenyuan Dai, and Qiang Yang. Network on network for tabular data classification in real-world applications, 2020.

[25] Guang Liu, Jie Yang, and Ledell Wu. Ptab: Using the pre-trained language model for modeling tabular data, 2022.

[26] Moe Kayali, Anton Lykov, Ilias Fountalis, Nikolaos Vasiloglou, Dan Olteanu, and Dan Suciu. Chorus: Foundation models for unified data discovery and exploration, 2024.

[27] Boris van Breugel, Jonathan Crabbé, Rob Davis, and Mihaela van der Schaar. Latable: Towards large tabular models, 2024.

[28] Yi He, Ke Xu, Shixiong Cao, Yang Shi, Qing Chen, and Nan Cao. Leveraging foundation models for crafting narrative visualization: A survey, 2025.

[29] Xianjie Wu, Jian Yang, Linzheng Chai, Ge Zhang, Jiaheng Liu, Xinrun Du, Di Liang, Daixin Shu, Xianfu Cheng, Tianzhen Sun, Guanglin Niu, Tongliang Li, and Zhoujun Li. Tablebench: A comprehensive and complex benchmark for table question answering, 2024.

[30] Zhe Li, Xiangfei Qiu, Peng Chen, Yihang Wang, Hanyin Cheng, Yang Shu, Jilin Hu, Chenjuan Guo, Aoying Zhou, Qingsong Wen, Christian S. Jensen, and Bin Yang. Foundts: Comprehensive and unified benchmarking of foundation models for time series forecasting, 2024.

[31] Dilermando Queiroz, Anderson Carlos, Maíra Fatoretto, Luis Filipe Nakayama, André Anjos, and Lilian Berton. Does data-efficient generalization exacerbate bias in foundation models?, 2024.

[32] Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling tabular data using conditional gan. *Advances in neural information processing systems*, 32, 2019.

[33] Han Wu, Zilong Zhao, Lydia Y. Chen, and Aad van Moorsel. Federated learning for tabular data: Exploring potential risk to privacy, 2022.

[34] Mohamed Abdelaal, Christian Hammacher, and Harald Schoening. Rein: A comprehensive benchmark framework for data cleaning methods in ml pipelines, 2023.

[35] Jihyeon Hyeong, Jayoung Kim, Noseong Park, and Sushil Jajodia. An empirical study on the membership inference attack against tabular data synthesis models, 2022.

[36] Lecheng Zheng, Baoyu Jing, Zihao Li, Hanghang Tong, and Jingrui He. Heterogeneous contrastive learning for foundation models and beyond, 2024.

[37] Sadaf Khademi, Zohreh Hajiakhondi, Golnaz Vaseghi, Nizal Sarrafzadegan, and Arash Mohammadi. Fh-tabnet: Multi-class familial hypercholesterolemia detection via a multi-stage tabular deep learning, 2024.

AI-generated, for reference only.

[38] Liangyu Zha, Junlin Zhou, Liyao Li, Rui Wang, Qingyi Huang, Saisai Yang, Jing Yuan, Changbao Su, Xiang Li, Aofeng Su, Tao Zhang, Chen Zhou, Kaizhe Shou, Miao Wang, Wufang Zhu, Guoshan Lu, Chao Ye, Yali Ye, Wentao Ye, Yiming Zhang, Xinglong Deng, Jie Xu, Haobo Wang, Gang Chen, and Junbo Zhao. Tablegpt: Towards unifying tables, nature language and commands into one gpt, 2023.

[39] Yubo Wang, Hao Xin, and Lei Chen. Kglink: A column type annotation method that combines knowledge graph and pre-trained language model, 2024.

[40] Jiuhai Chen, Jonas Mueller, Vassilis N. Ioannidis, Soji Adeshina, Yangkun Wang, Tom Goldstein, and David Wipf. Does your graph need a confidence boost? convergent boosted smoothing on graphs with tabular node features, 2022.

[41] Christos Koutras, Jiani Zhang, Xiao Qin, Chuan Lei, Vasileios Ioannidis, Christos Faloutsos, George Karypis, and Asterios Katsifodimos. Omnimatch: Effective self-supervised any-join discovery in tabular data repositories, 2024.

[42] Shengwei Li, Zhiquan Lai, Yanqi Hao, Weijie Liu, Keshi Ge, Xiaoge Deng, Dongsheng Li, and Kai Lu. Automated tensor model parallelism with overlapped communication for efficient foundation model training, 2023.

[43] Songheng Zhang, Haotian Li, Huamin Qu, and Yong Wang. Adavis: Adaptive and explainable visualization recommendation for tabular data, 2023.

[44] Michal K. Grzeszczyk, Szymon Płotka, Beata Rebizant, Katarzyna Kosińska-Kaczyńska, Michał Lipa, Robert Brawura-Biskupski-Samaha, Przemysław Korzeniowski, Tomasz Trzciński, and Arkadiusz Sitek. Tabattention: Learning attention conditionally on tabular data, 2023.

[45] Weichen Huang. Multimodal contrastive learning and tabular attention for automated alzheimer's disease prediction, 2023.

[46] Tennison Liu, Jeroen Berrevoets, Zhaozhi Qian, and Mihaela van der Schaar. Learning representations without compositional assumptions, 2023.

[47] Damien Garreau and Ulrike von Luxburg. Looking deeper into tabular lime, 2022.

[48] Timur Sattarov, Dayananda Herurkar, and Jörn Hees. Explaining anomalies using denoising autoencoders for financial tabular data, 2022.

[49] Samuel Stocksieker, Denys Pommeret, and Arthur Charpentier. Data augmentation with variational autoencoder for imbalanced dataset, 2024.

[50] Shi Bin Hoo, Samuel Müller, David Salinas, and Frank Hutter. The tabular foundation model tabpfn outperforms specialized time series forecasting models based on simple features, 2025.

[51] Mythreyi Velmurugan, Chun Ouyang, Catarina Moreira, and Renuka Sindhgatta. Developing a fidelity evaluation approach for interpretable machine learning, 2021.

[52] Tony T. Wang, Igor Zablotchi, Nir Shavit, and Jonathan S. Rosenfeld. Cliff-learning, 2023.

[53] Wenxuan Yang, Weimin Tan, Yuqi Sun, and Bo Yan. A medical data-effective learning benchmark for highly efficient pre-training of foundation models, 2024.

[54] Sourya Basu, Pulkit Katdare, Prasanna Sattigeri, Vijil Chenthamarakshan, Katherine Driggs-Campbell, Payel Das, and Lav R. Varshney. Efficient equivariant transfer learning from pretrained models, 2023.

[55] Manjunath Mahendra, Chaithra Umesh, Saptarshi Bej, Kristian Schultz, and Olaf Wolkenhauer. Convex space learning for tabular synthetic data generation, 2025.

[56] Hsin-Yu Chang, Pei-Yu Chen, Tun-Hsiang Chou, Chang-Sheng Kao, Hsuan-Yun Yu, Yen-Ting Lin, and Yun-Nung Chen. A survey of data synthesis approaches, 2024.

[57] Lei Xu and Kalyan Veeramachaneni. Synthesizing tabular data using generative adversarial networks, 2018.

[58] Dung Daniel Ngo, Daniel Scott, Saheed Obitayo, Vamsi K. Potluru, and Manuela Veloso. Adaptive and robust watermark for generative tabular data, 2024.

[59] Xiaotong Shen, Yifei Liu, and Rex Shen. Boosting data analytics with synthetic volume expansion, 2024.

[60] Conor Hassan, Robert Salomone, and Kerrie Mengersen. Deep generative models, synthetic tabular data, and differential privacy: An overview and synthesis, 2023.

[61] Hariharan Manikandan, Yiding Jiang, and J Zico Kolter. Language models are weak learners, 2023.

[62] Ahmed Alaa and Bin Yu. Veridical data science for medical foundation models, 2024.

[63] Haoyi Xiong, Xuhong Li, Xiaofei Zhang, Jiamin Chen, Xinhao Sun, Yuchen Li, Zeyi Sun, and Mengnan Du. Towards explainable artificial intelligence (xai): A data mining perspective, 2024.

[64] Cheng-Te Li, Yu-Che Tsai, Chih-Yao Chen, and Jay Chiehen Liao. Graph neural networks for tabular data learning: A survey with taxonomy and directions, 2024.

[65] David Chaves-Fraga, Edna Ruckhaus, Freddy Priyatna, Maria-Esther Vidal, and Oscar Corcho. Enhancing virtual ontology based access over tabular data with morph-csv, 2021.

Disclaimer:

SurveyX is an AI-powered system designed to automate the generation of surveys. While it aims to produce high-quality, coherent, and comprehensive surveys with accurate citations, the final output is derived from the AI's synthesis of pre-processed materials, which may contain limitations or inaccuracies. As such, the generated content should not be used for academic publication or formal submissions and must be independently reviewed and verified. The developers of SurveyX do not assume responsibility for any errors or consequences arising from the use of the generated surveys.

AI-generated, for reference only.