2023 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, SEPT. 17–20, 2023, ROME, ITALY

HTGTM: HYBRID TEMPORAL-GRAPH TABULAR MODEL FOR COMPLEX MULTIMODAL TABULAR DATA PROCESSING

Ziwen Liu, Scott Orr, Josep Grau-Bove

University College London
z.liu.19@ucl.ac.uk

ABSTRACT

Understanding the behaviour of volunteers is an important research area in social science and management. This paper tackles the Volunteer Retention Prediction task of the IEEE MLSP 2023 Data Challenge, utilizing a dataset from the COVID-19 pandemic volunteer coordination in Shenzhen, China in 2020 - 2021, with the objective of forecasting volunteer retention in the next ten months. Our paper proposes Hybrid Temporal-Graph Tabular Model (HTGTM), a deep learning-based hybrid model designed to extract and analyze temporal and graph information within complex tabular data. In this data challenge, we compared our model, ensembled with XGBoost, against traditional machine learning methods and deep learning models that are specifically tailored for tabular data. Our method exhibited robust performance, validated by its 1st-ranking Root-Mean-Square-Error (RMSE) score of 76.36 in MLSP 2023 Data Challenge Kaggle private Leaderboard. This research sheds light on pertinent volunteer retention prediction tasks and highlights the incorporation of deep learning techniques in complex, multimodal tabular data processing tasks.

Index Terms— Volunteer Retention, Deep Learning, Multimodal Tabular Data, Ensemble Learning

1. INTRODUCTION

The rapid advancement of digital technologies and platforms has brought about a revolutionary change in how people come together to address community needs and challenges. In particular, the emergence of online crowdsourcing platforms has transformed the landscape of volunteering activities, enabling individuals to contribute to collective goals in a more organized and effective manner. A notable example of this transformative power was observed during the COVID-19 pandemic, where volunteer self-organization played a crucial role in the collective response [1].

Through the utilization of crowdsourcing platforms and decentralized efforts, volunteers have been able to effectively address the urgent needs of communities. This shift in volunteering dynamics has highlighted the importance of understanding the behavior and collaboration patterns of these volunteers [1]. This paper focuses on the Volunteer Retention Prediction task of the IEEE MLSP 2023 Data Challenge on Kaggle[1]. The dataset utilized in this challenge was collected from the "Anti-Pandemic Pioneer", a platform used for self-organized volunteer coordination during the COVID-19 pandemic in Shenzhen, China [1].

The dataset spans self-organized volunteer coordination histories from February 2020 to May 2021, encapsulating essential data points of the volunteering activities such as participant identity, task timing and task geo-locations. This rich, complex tabular dataset intertwines temporal information derived from the chronological order of the tasks, and graph information extrapolated from user past co-operation data, thereby representing a multi-faceted view of the volunteering activities.

Given the unique characteristics and complexities of the dataset, building an effective predictive model is not straightforward. Notably, while the current state-of-the-art predictive models, namely deep learning-based models, excel in various domains, they often struggle when dealing with tabular data directly [2]. On the other hand, integrating data from different modalities could be a challenge for traditional machine learning methods [3]. Thus, this task calls for in-depth, focused research.

This paper elucidates Hybrid Temporal-Graph Tabular Model (HTGTM), an hybrid approach we have devised for this data challenge that holds the record for the highest performance (RMSE of 76.36) on the Kaggle private Leaderboard of this task. The outcome from this research formulates a robust predictive model for volunteer retention, and thus provides insights for improving volunteer management on crowdsourcing platforms. Furthermore, this research carries broader implications by exploring the integration of deep learning models into complex tabular datasets, where multimodal information, such as temporal and graph data structures, are encompassed.

[1]https://kaggle.com/competitions/MLSP2023-volunteers-01

979-8-3503-2411-2/23/$31.00

2. RELATED WORK

2.1. Volunteer Retention

Understanding what makes volunteering initiatives successful is important due to its rising significance in promoting sustainability and benefiting society [4]. Numerous studies have indicated that volunteers' perspectives on their experiences, such as a sense of fulfillment, satisfaction, and contentment, strongly predict their future retention [5, 6]. In addition to these personal perception variables, volunteers' commitment, as indicated by their past volunteering records, and their interpersonal relationships within the volunteering community have also been found to be related to their continued engagement [4, 5, 6]. Since this dataset does not include variables related to personal perception, volunteers' past participation records and their engagement with the community, such as records of collaboration with others, are important factors to consider when modeling their future retention.

2.2. Deep Learning for Tabular Data

Deep learning models have proven to be highly effective in various domains, including computer vision and natural language processing, thanks to their powerful automatic feature extraction capabilities. However, when it comes to modeling tabular data, traditional machine learning methods such as gradient-boosted decision trees (GBDT) have continued to dominate and consistently outperform deep learning approaches [2]. Recent advancements in neural network architecture research have led to the development of deep learning methods specifically designed for tabular data, which claim to offer superior performance compared to tree-based methods [7, 8, 9, 10]. However, the question of whether deep learning models outperform traditional methods on tabular data is still a topic of controversy due to the absence of standard benchmark for holistically comparing different methods' performance on tabular data processing tasks [2].

Despite the ongoing debate, certain neural network architectures have proven their superior capabilities in handling non-tabular data [11]. These data types, such as time-series and graph data, frequently appear either explicitly or implicitly in or alongside tabular data, rendering traditional methods less suitable for their analysis. Hence, it remains essential to consider the incorporation of deep learning in such complex multimodal tabular data processing.

3. METHODOLOGY

3.1. Data Exploration

The MLSP 2023 Volunteer Retention Prediction dataset contains approximately 2.4 million entries of volunteering records, with user ID, task ID, task locations (geo-coordinates and location names), task timing (task start and end time) and an indicator of whether the user of the entry being task organizer. There are 47,617 unique users in total among the 2.4 million entires which have been evenly split into the training ant test set. The task ID column embeds user's past co-operation history and thus inherently signifies the underlying network structure among these 47,617 unique users. Moreover, the temporal and spatial information derived from the task time and locations respectively, combines to form individual temporal-spatial trajectories for each user.

Figure 1 (a) visually illustrates the structure of the dataset, taking into account both time-series and graph information. The visualization reveals that the dataset incorporates data modalities beyond the tabular form: From a temporal perspective, each unique user is associated with an unequal length time-series spreadsheet that records his/her temporal-spatial trajectory of past volunteering activities. This temporal dimension captures the sequence of events and locations for each user's volunteering engagements. From a graph perspective, the relationships between users are represented as a network structure, where each user is represented as a node, and the number of past co-operations between users is depicted as edge links connecting the corresponding nodes. Additionally, the frequency of co-operation between users in the past indicates the strength of the edge link.

To address the task objective, a straightforward strategy involves incorporating graph information into a tabular format and aggregating the temporal dimension of the dataset. This creates user-level aggregated task statistics for each user and this task becomes a regression task on a classical tabular dataset. By doing so, conventional techniques for working with tabular data can be applied. However, this approach poses the risk of losing detailed information when aggregating high-frequency time-series data. Furthermore, seamlessly integrating connections in graph data into useful features in tabular data is challenging. To tackle these challenges, our paper proposes a hybrid approach: (1) as previously mentioned, we aggregate the temporal dimension and discard the graph information to create a standard tabular dataset, on which we perform feature engineering and build a tabular model; (2) in parallel, we feed the time-series data and graph data into separate models, namely a time-series model and a Graphical Neural Network (GNN) model. The three models will compose a hybrid model and be trained jointly. This will be explained in greater detail in Section 3.3.

3.2. Data Preparation and Feature Engineering

This section describes the feature engineering we conducted on the aggregated tabular dataset. Following the literature in Section 2.1, volunteers' commitment (as represented by past volunteering activities in this task) is a crucial indicator for their future retention. As a result, we artificially created several features to extract useful information towards this direction.

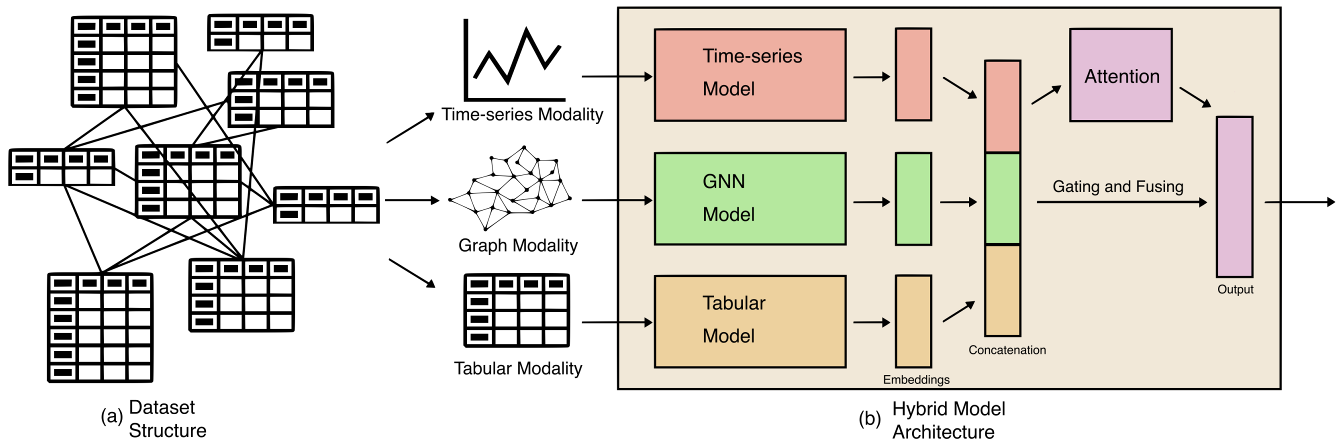

Fig. 1. (a) Visualization of the dataset structure incorporating time-series and graph information. Each user has a corresponding time-series volunteering activity recordings and the connections between users can be inferred from their past co-operations in volunteering tasks. (b) Hybrid model architecture consisting of three sub-models with specialized architectures for distinct data types. Each sub-model processes the input data into embeddings. The resulting embeddings are then gated using an attention module and fused to produce the final output. More details regarding each component are introduced in Section 3.3

Specifically, we split 2020 - 2021 into 4 quarters and counted the total number of past volunteering tasks and the number of times each user served as an organizer for these tasks in each quarter. To delve deeper into the data, we calculated the total time span from February 2020 to May 2021 in hours and determined the percentage of each user's participation within this duration. Additionally, we categorized each day within the time span as a working day, weekend, or public holiday, and further divided each day into three distinct time periods (20:00 - 04:00, 04:00 - 12:00, and 12:00 - 20:00). By doing so, we were able to quantify the number of volunteering tasks participated in by each user during specific days and time periods. These newly created features may offer insights into a volunteer's willingness to engage in both regular and atypical dates and times, thereby providing a more comprehensive measure of their commitment.

Afterward, we computed the centroid of each user's geo-coordinates to serve as a representation of their previous volunteering locations. Additionally, we derived new features by calculating the variation and distance traveled based on these geo-coordinates. These features may offer insights into whether a volunteer is inclined to engage solely in tasks within their local vicinity or if they demonstrate a willingness to participate in activities that are farther away from their usual areas. As a result, these features provide an indication of the volunteer's commitment level and their readiness to venture beyond their routine regions.

Lastly, since the platform is closely tied to the COVID-19 pandemic in Shenzhen, we incorporated external data on the daily COVID-19 cases in Guangdong province, China (where Shenzhen is located). We then analyzed the correlation between the 7-day moving-averaged COVID-19 cases (local and imported) and the user's participation throughout the entire time span. This analysis aimed to shed light on whether volunteers predominantly concentrated their efforts on COVID-19 related tasks or demonstrated a willingness to engage in activities unrelated to the pandemic. This information may provide insights into the volunteers' intentions and potential for future retention.

In addition to the preceding feature engineering, we also implemented a reweighting scheme based on the future retention distribution of volunteers. Specifically, we have observed that this distribution displays characteristics of a power-law, suggesting that outliers in the right-tail should receive more attention [12]. As a result, we tailored our approach to give these outliers greater consideration by employing a reweighting technique: We fitted the target variable (future retention), denoted by y, to a Pareto distribution and obtained the probability density function $f_{\hat{\theta}}(y)$ of the distribution, where $\hat{\theta}$ denotes the estimated parameters. Based on $f_{\hat{\theta}}(y)$, we computed a weight w for each user in the training set according to the equation:

$$w_{u_t} = \frac{\ln(C + \frac{1}{f_{\hat{\theta}}(y)})}{\frac{1}{N_t}\sum_{y\in\mathcal{D}t}\ln(C + \frac{1}{f_{\hat{\theta}}(y)})}$$

In the equation, u_t denotes the user ID in the training set, $\mathcal{D}t$ the training set, and Nt is the total count of users in the training set. C is used as a smoothing constant (we set $C = 100$ heuristically). These weights assign greater values to data points that are less likely according to the fitted Pareto distribution, thus enabling the model to focus more on the volunteers who are likely to have high retention in the future.

3.3. Proposed Model Architecture

This section presents the proposed model architecture for the joint processing of time-series, graph, and aggregated tabular data extracted from the dataset. As shown in Figure 1 (b), our approach is a hybrid model incorporating three submodels with different architectures: a time-series model, a Graph Neural Network (GNN) model and a tabular model. Each model receives a distinct type of inputs and generates corresponding embeddings. These embeddings are then concatenated togater (denoted as E_{cat}).

Subsequently, as the three forms of the input data are originated from the same dataset, there is a potential redundancy of overlapped information from each model's output embedding. To reduce this redundancy, we employed an **attention module**: The attention module takes the concatenated embeddings E_{cat} as input and produces gating scores G, ranging from 0 to 1, for each dimension of the embeddings. The gating scores determine the relevance or importance of each dimension in E_{cat}. Next, we calculate the dot product between the gating scores and the concatenated embeddings as gated embeddings $E_{gated} = E_{cat} \cdot G$, which is then fused and projected by a linear layer. Finally, an output layer is employed as the final layer to generate predictions.

By combining the strengths of the time-series, GNN, and tabular models, our proposed architecture leverages temporal information, graph structure, and aggregated statistics to improve the overall predictive performance. The attention module enables the model to dynamically attend to the most informative dimensions in the concatenated embeddings, facilitating effective feature fusion and enhancing prediction accuracy. We hereby refer to our hybrid model as the Hybrid Temporal-Graph Tabular Model (HTGTM). We then introduce the time-series model, GNN model and tabular model in HTGTM respectively as follows:

Time-series Model: Since the time-series data for each user has unevenly spaced intervals between volunteering tasks, traditional recurrent neural network architectures like GRU [13] and LSTM [14] are not suitable in this case as they sequentially process each timestamp which implicitly assumes evenly spaced time-series data. To address this issue, we adopted a decoder-only transformer architecture. Specifically, we chose the day as the granularity for processing the time-series sequence, and initialized trainable embeddings for each day from February 2020 to May 2021 serving as positional embedding for each token in the sequence. Each participation entry for a given date, including geo-coordinates, the number of tasks on the date, and task length, was treated as a token embedding in the sequence. Following the standard practice in transformer models, the positional embedding and token embedding were mixed by addition before self-attention operation. During self-attention operation, we masked out future days to ensure the model operates in a unidirectional manner. By adopting this approach, we effectively handle uneven intervals between volunteering activities in the time-series data while preserving the time-series nature of the information.

Graph Neural Network Model: Our GNN model utilizes a graph convolutional network architecture, comprising three graph convolutional layers. To capture the connections in the graph that is represented by the inter-relations among the community's 47,617 unique users, we construct an affinity matrix based on past instances of co-operations between pairs of users. The weight assigned to each edge corresponds to the number of co-operations observed. In order to initialize node features in the model, we leverage the tabular data features available for each user.

Tabular Model: We applied TabNet [8] as the tabular model which receives aggregated tabular data. TabNet is a transformed-based model for tabular data that uses an encoder-decoder architecture with sparse feature selection mechanisms to learn feature interactions. We took its supervised learning version and followed default settings and hyper-parameters given in the original paper [8]. We only modified its final layer to produce embeddings instead of predictions.

Lastly, we acknowledge the inherent limitations of neural networks, such as their tendency to produce over-smooth solutions and their susceptibility to the negative impact of uninformative features [11]. These drawbacks can potentially result in degraded performance for this regression task. Thus, as recommended by [2], we took an ensemble approach to make the final predictions. To be more specific, we concatenate the features in tabular data with the prediction from HTGTM, and fed them into an XGBoost regressor [15] for further training, and the output will be the final predictions from our method. This ensemble approach allows us to leverage the feature extraction capabilities of deep learning models while incorporating the robustness of tree-based methods. We refer to this ensemble as XGBoost+HTGTM.

4. EXPERIMENTAL SETUP AND RESULTS

4.1. Experimental Setup

Our HTGTM model was trained using the Adam optimizer [16] with a learning rate of 10^{-3}, a batch size of 32, and for 10 epochs. The mean squared error loss was used, and the proposed reweighting scheme was introduced after the initial 5 epochs. For the XGBoost+HTGTM ensemble, we optimized the hyperparameters of the XGBoost regressor using random grid search based on the results of 3-fold cross-validation. To ensure speed, we utilized a sub-sample of 10,000 randomly selected from the training set during the random grid search.

In this study, we evaluated our proposed model utilizing the Root-Mean-Square-Error (RMSE) metric. We provide the evaluation results from three sources: the score obtained from a holdout validation set, consisting of 33% of randomly selected data from the training set (carefully partitioned to maintain comparable data distribution in both the training and validation sets); the Kaggle public Leaderboard score (computed on 33% of the test set, which is visible to all participants throughout the competition); and the Kaggle private Leaderboard score (calculated on the remaining 66% of the test set, which was kept hidden during the competition and disclosed only at the end to determine the final standings).

Given the restriction of two submissions per day for the competition, exhaustive model comparisons were impracticable. However, leveraging our expertise, we chose the following models as benchmarks:

XGBoost [15]: A GBDT algorithm known for its speed and performance, especially on tabular data processing.

LightGBM [17]: Another GBDT algorithm that grows trees leaf-wise, resulting in smaller and faster models while maintaining similar level of performance to XGBoost.

CatBoost [18]: Another GBDT method with a special emphasis on categorical features.

FT-Transformer [9]: A simple adaptation of the Transformer architecture for the tabular domain that transforms features (categorical and numerical) to tokens and runs a stack of layers on the tokens.

TabTransformer [7]: Another transformer-based model for tabular data that provides both supervised and unsupervised learning functions. Here we only use its supervised learning version.

Tree-based methods were fitted to tabular data exclusively with the same hyperparameter searching strategy used for XGBoost+HTGTM (as described above). Additionally, to ensure a fair comparison, we also report the performance of an ensemble version of each deep learning-based method, following the same procedure as used in XGBoost+HTGTM.

Finally, to further investigate whether our hybrid model is dominated by a single sub-model within its architecture, we performed an ablation study by comparing HTGTM to each sub-model within HTGTM, both individually and as an ensemble with XGBoost. Each sub-model is trained with the same training and ensemble procedure as introduced above. We only modified their output layer to produce predictions instead of embeddings.

4.2. Results

Table 1 presents the results of our proposed method with other competitive models listed in Section 4.1. It is observed that our proposed ensemble method, XGBoost+HTGTM, achieved lowest RMSE score. Also, when compared with other deep learning-based models, HTGTM has the best performance, suggesting that incorporating temporal and graph information from the dataset enhances predictive power. It is also observed that tree-based models consistently outperform deep learning-based models on this task, echoing the advantage of tree-based models on tabular data tasks.

		Kaggle	
	Val Set	Public	Private
TabTransformer	103.22	110.25	104.50
FT-Transformer	104.59	111.47	105.44
GNN Model-only (GNN†)	88.24	97.73	91.85
Time-series Model-only (TS†)	86.49	93.66	89.34
Tabular Model-only (TAB†)	111.04	100.60	94.14
HTGTM	**77.38**	**90.12**	**81.26**
LightGBM	71.93	84.91	78.80
CatBoost	70.41	82.91	77.98
XGBoost	69.47	83.61	78.76
XGBoost+TabTransformer	81.61	83.84	80.65
XGBoost+FT-Transformer	87.51	86.34	80.75
XGBoost+GNN†	69.04	83.88	78.52
XGBoost+TS†	68.96	85.19	79.15
XGBoost+TAB†	70.12	85.48	79.17
XGBoost+GNN†+TS†+TAB†	69.02	84.35	77.53
XGBoost+HTGTM	**66.27**	**79.09**	**76.36**

Table 1. Model comparisons with baseline models and ablated versions, i.e. with individual sub-model only, superscribed by †, on holdout validation (Val) set, Kaggle public and private Leaderboard. Results are reported in RMSE.

Additionally, it is worth noting that not all ensembles lead to a reduction in prediction error based on Table 1. The ensembles of XGBoost+TabTransformer and XGBoost+FT-Transformer showed decreased performance compared to XGBoost alone. This suggests that the output produced by TabTransformer and FT-Transformer did not contribute useful additional information to improve XGBoost's learning from the existing features and highlights the improtance of incorporating graph and time-series information.

The results of the ablation study are also presented in Table 1, where models superscribed by † are ablated versions, i.e. sub-models within HTGTM. It is observed from the table that HTGTM outperforms each individual sub-model within its hybrid architecture. Similarly, our ensembled version, XGBoost+HTGTM, surpasses each ensembled version of sub-models. This outcome suggests the benefit of multi-modality, joint training and the attention module's gating and fusing mechanism within the hybrid model. Overall, the ablation study confirms that no single sub-model is dominating the performance of the hybrid model and thus shows the robustness of HTGTM and the efficacy of its integrated approach.

Finally, due to the inherent randomness in our proposed XGBoost+HTGTM, such as model parameters initialization, dropout layers and batch training, we re-ran it five times on these datasets after the competition ends to quantify the variability of our model's predictions. We obtained the following results (reported as mean $\pm$ standard errors): 67.03$\pm$0.59 on the heldout validation set, 80.74$\pm$0.76 on the Public Leaderboard and 75.20$\pm$0.98 on the Private Leaderboard.

5. CONCLUSION

In this paper, we introduced HTGTM, a deep learning hybrid model designed to address the Volunteer Retention Prediction task in the IEEE MLSP 2023 Data Challenge. The proposed method can effectively integrate various types of data embedded in a complex multimodal tabular dataset. Our method, when ensembled with XGBoost, showed state-of-the-art performance as measured by RMSE, which is demonstrated by its first place in the Kaggle private Leaderboard against other strong baselines and competitors in the data challenge. This research not only provides a promising approach for volunteer retention prediction but also makes a broader contribution to the field by showcasing the effective integration of deep learning techniques into complex tabular data processing. Finally, in our current work, interactions between the sub-models occur solely after the concatenation of their produced embeddings, making it a loose hybrid model. Going forward, we aim to explore directions that foster more interactions between the sub-models in early stages to further enhance the model's multimodal ability.

6. REFERENCES

[1] Anping Zhang, Ke Zhang, Wanda Li, Yue Wang, Yang Li, and Lin Zhang, "Optimising self-organised volunteer efforts in response to the covid-19 pandemic," *Humanities and Social Sciences Communications*, vol. 9, no. 1, 2022.

[2] Ravid Shwartz-Ziv and Amitai Armon, "Tabular data: Deep learning is not all you need," *Information Fusion*, vol. 81, pp. 84–90, 2022.

[3] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 2, pp. 423–443, 2018.

[4] Steven Bauer and Dongkuk Lim, "Effect of communication practices on volunteer organization identification and retention," *Sustainability*, vol. 11, no. 9, pp. 2467, 2019.

[5] Vivien W Forner, Djurre Holtrop, Darja Kragt, and Anya Johnson, "The great (volunteer) resignation: An evidence-based strategy for retaining volunteers," 2022.

[6] Beth Gazley, "Predicting a volunteer's future intentions in professional associations: A test of the penner model," *Nonprofit and Voluntary Sector Quarterly*, vol. 42, no. 6, pp. 1245–1267, 2013.

[7] Xin Huang, Ashish Khetan, Milan Cvitkovic, and Zohar Karnin, "Tabtransformer: Tabular data modeling using contextual embeddings," *arXiv preprint arXiv:2012.06678*, 2020.

[8] Sercan Ö Arik and Tomas Pfister, "Tabnet: Attentive interpretable tabular learning," in *Proceedings of the AAAI conference on artificial intelligence*, 2021, vol. 35, pp. 6679–6687.

[9] Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko, "Revisiting deep learning models for tabular data," *Advances in Neural Information Processing Systems*, vol. 34, pp. 18932–18943, 2021.

[10] Gowthami Somepalli, Micah Goldblum, Avi Schwarzschild, C Bayan Bruss, and Tom Goldstein, "Saint: Improved neural networks for tabular data via row attention and contrastive pre-training," *arXiv preprint arXiv:2106.01342*, 2021.

[11] Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux, "Why do tree-based models still outperform deep learning on typical tabular data?," *Advances in Neural Information Processing Systems*, vol. 35, pp. 507–520, 2022.

[12] GC Crawford, "Skewed opportunities: How the distribution of entrepreneurial inputs and outcomes reconceptualizes a research domain," *Proceedings-Academy of Management*, 2018.

[13] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[14] Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

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

[16] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[17] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu, "Lightgbm: A highly efficient gradient boosting decision tree," *Advances in neural information processing systems*, vol. 30, 2017.

[18] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin, "Catboost: unbiased boosting with categorical features," *Advances in neural information processing systems*, vol. 31, 2018.