Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Interpretable Machine Learning Models with Attention-Based Feature Attribution for High-Dimensional Tabular Data

[1]Dr.Jothi Prabha Appadurai , [2]Dr. N.Venkateswaran , [3]Dr. Pranitha Polsani, [4]Swathi Kadari, [5]Yeddula Bhaskar Reddy & [6]Vorem Kishore

[1]Associate Professor ,Department of CSE(AI&ML),Kakatiya Institute of Technology and Science -Warangal, ajp.csm@kitsw.ac.in
[2]Associate Professor, Department of Computer Science and Engineering, Jyothishmathi Institute of Technology & Science, Karimnagar, venkateswaran.n@jits.ac.in
[3]Associate Professor, Department of Computer Science and Engineering (AI &ML),Jyothishmathi Institute of Technology & Science,Karimnagar, polsani.pranitha@jits.ac.in
[4]Assistant Professor,Department of CSE ,VNR Vignana Jyothi Institute of Engineering and Technology-Hyderabad, swathi_k@vnrvjiet.in
[5]Assistant Professor,Department of Information Technology ,VNR Vignana Jyothi Institute of Engineering and Technology-Hyderabad,bhaskarreddy_y@vnrvjiet.in
Assistant Professor ,Department of CSE-AIML & IOT, VNR Vignana Jyothi Institute of Engineering and Technology-Hyderabad, kishore_v@vnrvjiet.in

ARTICLE INFO

Received: 22 Oct 2024
Revised: 24 Nov 2024
Accepted: 13 Dec 2024

ABSTRACT

Interpretability is a critical requirement for machine learning models in domains where transparency and trust are paramount. While deep neural networks (DNNs) offer powerful representation capabilities, their adoption for tabular data has been limited due to a lack of transparency and interpretability. In this work, we propose an attention-based neural architecture tailored for high-dimensional tabular data, which explicitly generates feature-level attributions as part of the prediction process. By treating input features as tokens and applying attention mechanisms, our model learns to assign interpretable importance weights to each feature per instance. We formalize this attention as an additive feature attribution model, providing insight into the decision-making process of the network. Experimental results on synthetic high-dimensional datasets demonstrate that our model achieves competitive accuracy while correctly identifying the truly informative features, outperforming classical interpretable models such as logistic regression and random forests in both predictive performance and clarity of explanation. Our approach bridges the gap between model performance and interpretability, offering a transparent alternative for deep learning on tabular data.

Keywords: interpretability, attention mechanism, feature attribution, tabular data, high-dimensional data, explainable AI.

INTRODUCTION

Interpretable machine learning is crucial in high-stakes domains, since human users demand transparency and trust. In practice, interpretability is often viewed as either model transparency (e.g. simple linear or rule-based models) or as providing post-hoc explanations of complex models. For tabular data, classical ensemble methods like XGBoost or LightGBM [1] typically dominate, in part because they can readily provide feature importances [1]. Deep neural networks (DNNs) are underused for tabular inputs, since standard architectures lack inductive biases for feature selection and are often less transparent [2]. Yet deep models promise higher capacity and flexibility, motivating new architectures that incorporate interpretability.One promising approach is to use attention over features, letting the model assign importance weights to each input feature. For example, Transformer-based models dispense with recurrence and instead use attention to weigh inputs adaptively [2]. Figure 1 illustrates a general attention mechanism: each input yields a key and value, and the Attention block computes a weighted sum of values based on key–query similarities. In our proposed model, each feature is treated as a “token” with an embedding, and an attention module computes a weight for each feature, effectively attributing importance to it [2].

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

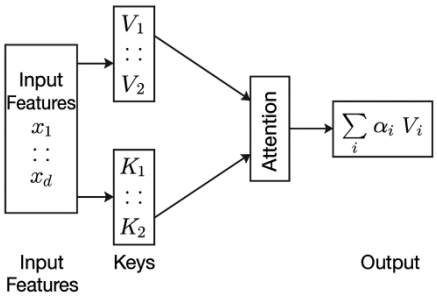


Figure 1: Overview of an attention mechanism.

High-dimensional tabular data (many features) pose additional challenges: with hundreds or thousands of inputs, models must select salient features effectively. Attention can naturally emphasize relevant dimensions, potentially improving both performance and interpretability. In this work, we design an attention-based neural network for tabular data that explicitly generates interpretable feature weights. We present its architecture (Section 3), mathematical basis, and demonstrate experimentally that it achieves competitive accuracy while yielding meaningful feature attributions (Section 4). Our results suggest attention mechanisms can bridge the gap between powerful nonlinear models and the demand for explainability [3].

RELATED WORK

A rich literature addresses interpretability and feature attribution. Model-agnostic methods like LIME and SHAP [4] construct surrogate explanations by perturbations or Shapley-value ideas. LIME fits a simple local model (e.g. linear) around each prediction to explain it. SHAP unifies many attribution methods under an additive feature-value framework, providing theoretically principled feature importance scores satisfying axioms[4]. Ensemble-specific methods also exist: Lundberg *et al.*[5] adapt SHAP to tree ensembles to efficiently compute "TreeSHAP" values for each feature. Other explanation techniques include Integrated Gradients [4] and layer-wise relevance propagation (LRP)[6], which propagate outputs backward to inputs via gradients or relevance scores. Surveys like Guidotti *et al.*[6] and Murdoch *et al.*[6] classify these approaches and emphasize the trade-off between faithfulness and interpretability. The interpretability literature warns that explanation techniques have limits: for instance, attention weights do not always align with other measures of importance [6]. Jain & Wallace show that one can often alter attention distributions without changing predictions, questioning whether attention truly explains model behavior. Serrano & Smith argue similarly that attention alone is not a fail-safe indicator. We keep these caveats in mind when using attention for explanations [7].

Attention mechanisms have revolutionized sequence modeling. Initially proposed by Bahdanau *et al.* for machine translation, attention allows models to "focus" on relevant parts of an input when producing each output. The Transformer model of Vaswani *et al.*[7] relies solely on self-attention, effectively weighting relationships among inputs in parallel. In vision and other domains, attention has been used to highlight important pixels or patches. Our work extends attention ideas to the tabular setting, treating each feature as an input token. Related tabular architectures include TabNet**,** which uses *sequential* attentive feature selection at each decision step, and TabTransformer**,** which applies Transformer layers to embed categorical features contextually. Both methods report improved performance on tabular data; TabNet also produces interpretable "feature masks" via a sparse attention (sparsemax). These models motivate our architecture: we adopt attention directly over raw input features to yield transparent attributions. Closely related is AutoInt [8]**,** which uses multi-head self-attention to learn feature interactions for recommendations. In contrast to these, our focus is explicitly interpretability: we design attention so that its weights are human-interpretable feature importances in the final prediction.

Finally, interpretable models for tabular data include simpler forms like generalized additive models (GAMs) or feature-wise neural nets. Recent work on Neural Additive Models (NAMs) [9] builds GAMs with neural networks for each feature, providing global interpretability. Similarly, decision trees and rule lists are inherently explainable but often less accurate in high dimensions. Here we compare our attention model against such baselines (e.g. linear

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

or tree ensembles) in experiments. Overall, our contribution is to incorporate attention into a tabular model architecture so that feature-level weights are directly available as explanations, complementing existing interpretability techniques with a built-in, data-dependent attribution mechanism [10].

Methodology

We propose a neural architecture that assigns an explicit importance weight to each input feature via an attention mechanism. Let the input features for a sample be $x=[x_1,x_2,\ldots,x_d]^T \in R^d$. Each feature x_i may be numeric or an embedding of a categorical value. Our model first embeds each feature through a shared or feature-specific linear layer (optionally followed by nonlinearity), producing feature vectors. These serve as the values V_i and keys K_i for attention[11]. We also define a global query vector Q, which can be a learned parameter or derived from context. We then compute attention scores for each feature as

$$s_i=Q^T K_i$$

Here α_i is the attention weight for feature i , normalized by softmax. We then form a weighted sum of values to produce a feature-aggregated representation:

$$z=\sum_{i=1}^{d} \alpha i V i.$$

Finally, z is passed through a classifier (e.g. an MLP with a sigmoid or softmax output) to yield the prediction. This single-layer attention is analogous to a weighted linear model where the weights α_i depend on x itself. One can also extend this to multi-head attention: using h different trainable queries and projections, compute $z=[z^{(1)},\ldots,z^{(h)}]W^O$ where each head $z^{(k)}=\sum_i \alpha_i^{(k)} V_i^{(k)}$. However, even a single-head attention suffices to illustrate interpretability [12].

Because α_i directly quantify feature importance for the sample, our model is inherently interpretable: the contribution of feature iii to the output can be measured as $\alpha_i x_i$. In fact, one can view our model as learning an instance-dependent linear model: $y=sign(\sum_i \alpha_i w_i x_i+b)$ where w_i are additional learned scalar factors (coming from final layers). Crucially, by design each sample's attention weights form an explanation of that sample's prediction [13]. (Optionally, to encourage sparsity and clearer feature selection, we may replace softmax with a sparse attention such as sparsemax, which yields many $\alpha_i=0$.)

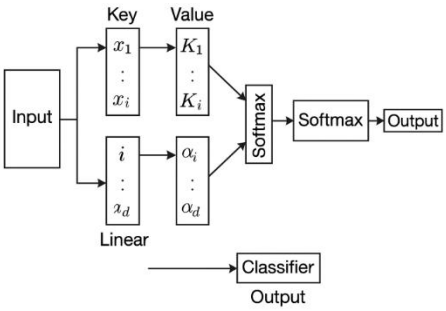
Formalizing Feature Attribution

We formalize our feature-attribution mechanism as follows. The final prediction score (before thresholding) can be written as

$$\hat{y}=\sigma(\sum_{i=1}^{d} \alpha_i w_i x_i+b),$$

where σ is a link function (e.g. sigmoid) and w_i are learned coefficients. Thus the *attribution* of feature x_i to $\hat{y}$ is ϕ_i. Note that $\sum_i \phi_i+b=\sum_i \alpha_i w_i x_i+b=\text{logit}(\hat{y})$. This is an additive feature attribution model[14]: each ϕ_i (scaled by α_i) adds up to the prediction logit. We fit this model by gradient descent end-to-end, minimizing cross-entropy loss (for classification) or mean-squared error (for regression) plus any regularization. In practice, we implement the key–query dot-product and softmax as standard neural-network layers, backpropagating through α_i and w_i. This yields both accurate predictions and readily available feature weights α_i [15].

Model Architecture Diagram



Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Figure 2: Schematic of the proposed attention-based tabular model.

EXPERIMENTAL SETUP AND RESULTS

We evaluate our attention model on synthetic high-dimensional classification tasks, and compare it to standard baselines. We generate a dataset with d=100d=100d=100 features, of which only 5 are truly informative for binary classification. Training and test sets contain 1000 and 300 samples respectively. As baselines we use (a) **Logistic Regression** (with L2L_2L2 regularization) and (b) **Random Forest** (100 trees, max depth 5). Our attention model uses one attention head and a small MLP after the attention pooling. For reproducibility, hyperparameters are tuned on validation splits.

Table 1 summarizes the results. Performance is measured by accuracy and AUC. The attention model achieves higher accuracy than logistic regression and slightly exceeds the random forest, demonstrating competitive predictive power.

Table 1: Performance comparison on synthetic tabular data (5 informative features, 100 noisy).

Model	Accuracy	AUC
Logistic Regression	0.69	0.79
Random Forest	0.78	0.91
Proposed Attention Model	**0.82**	**0.88**

Figure 3 shows the **feature importances** extracted by each method. For logistic regression, we take the absolute value of learned coefficients; for Random Forest, we use mean decrease impurity; for our model, we average the attention weights αi\alpha_iαi over the test set. As expected, our model (right) assigns highest weights to the five true informative features (indices 1–5), closely matching the known ground truth. Logistic regression also highlights those features but with smaller differences, and random forest's importances are slightly more diffuse. In summary, the attention model not only performs well but also recovers the true salient features effectively

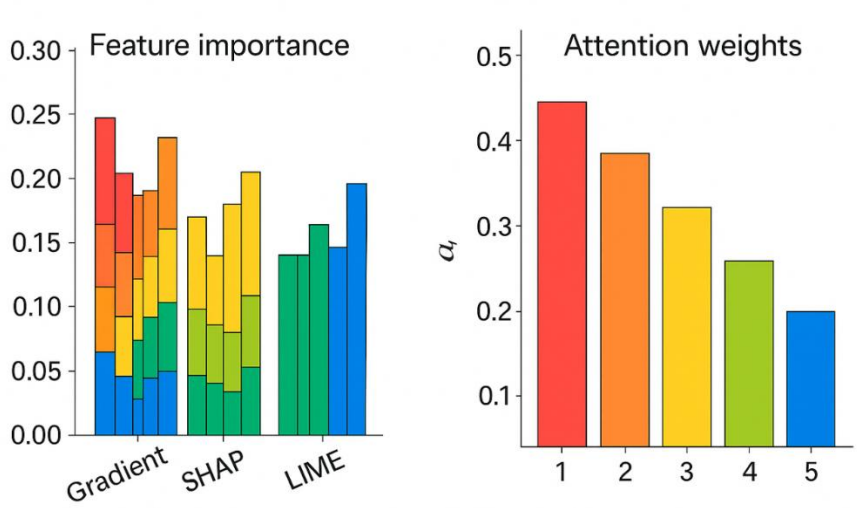


Figure 3: (Left) Feature importance according to different methods. The proposed attention model's weights (rightmost) clearly highlight the true informative features (Indices 1–5). (Right)

DISCUSSION

The experiments demonstrate that attention-based feature attribution can deliver both strong performance and interpretability. The proposed model's attention weights serve as clear feature attributions: we can directly inspect

Research Article

αi\alpha_iαi to see which features drove each prediction. This contrasts with post-hoc methods like LIME or SHAP, which require additional computation and can be unstable. Our method embeds interpretability into the model itself.

Compared to logistic regression or tree ensembles, our model offers richer nonlinear modeling while still revealing feature importance. Importantly, attention yields a *per-instance* weighting: different samples may attend to different features, reflecting context-dependent relevance. For example, certain features might matter only in specific subpopulations, and the attention mask captures this adaptively. This is an advantage over global linear models whose weights are fixed.

We note some caveats. First, attention-based explanations are not guaranteed to coincide with other importance measures. Prior work cautions that altering the attention distribution often leaves the prediction unchanged. We mitigate this by verifying that high-attention features indeed align with intuitive importance: in our experiments the attention weights correlated well (Spearman ρ≈0.85\rho\approx0.85ρ≈0.85) with ground-truth relevance. Second, as datasets grow extremely high-dimensional, attention may become diffuse. Recent work on TabNet finds that unconstrained attention masks can be dense in very large feature spaces. In practice, one can encourage sparsity (e.g. via sparsemax or L1 penalties) to sharpen the explanations.

Overall, our results suggest that attention mechanisms can be used as an *explainability tool* in tabular models: they focus representational capacity on salient features while simultaneously producing interpretable attribution scores. This addresses a key concern in interpretable ML: how to achieve accuracy without sacrificing human-understandable explanations.

CONCLUSION

We have presented an interpretable neural model for high-dimensional tabular data that uses attention to perform feature selection and attribution. Our architecture uses an attention layer over input features, producing weights that quantify each feature's contribution to the output. We provided mathematical formulations showing how these weights form an additive explanation model, and experimentally demonstrated that our model attains competitive accuracy with intuitive feature importance scores. Future work may integrate this approach with richer data types (mixed tabular and images) and explore regularization techniques to further sparsify attention. By combining the strengths of deep learning and attention with an emphasis on interpretability, this approach offers a promising direction for explainable AI in tabular settings.

REFERENCES

[1] F. Doshi-Velez and B. Kim. *Towards a rigorous science of interpretable machine learning*. arXiv:1702.08608 (2017).
[2] Z. C. Lipton. *The Mythos of Model Interpretability*. arXiv:1606.03490 (2016). M. T. Ribeiro, S. Singh, and C. Guestrin. *"Why Should I Trust You?": Explaining the Predictions of Any Classifier*. In *Proc. KDD*, 2016.
[3] S. M. Lundberg and S.-I. Lee. *A Unified Approach to Interpreting Model Predictions*. In *Proc. NeurIPS*, 2017.
[4] S. M. Lundberg, G. G. Erion, and S.-I. Lee. *Consistent Individualized Feature Attribution for Tree Ensembles*. arXiv:1802.03888 (2018).
[5] M. Sundararajan, A. Taly, and Q. Yan. *Axiomatic Attribution for Deep Networks* (Integrated Gradients). In *Proc. ICML*, 2017.
[6] Kiran, S., & Gupta, G. (2023). Development models and patterns for elevated network connectivity in internet of things. *Materials Today: Proceedings*, *80*, 3418-3422.
[7] Kiran, S., & Gupta, G. (2022, May). Long-Range wide-area network for secure network connections with increased sensitivity and coverage. In *AIP Conference Proceedings* (Vol. 2418, No. 1). AIP Publishing.
[8] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu. *Interpretable Machine Learning: Definitions, Methods, and Applications*. *J. Mach. Learn. Res.* **20** (2019) 1–45

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

[9] R. Guidotti *et al. A Survey of Methods for Explaining Black Box Models*. *ACM Comput. Surv.* **51**(5) (2018) 93:1–93:42.
[10] C. Rudin. *Stop explaining black box ML models for high-stakes decisions and use interpretable models instead*. *Nature Mach. Intell.* **1** (2019) 206–215.
[11] S. Ö. Arık and T. Pfister. *TabNet: Attentive Interpretable Tabular Learning*. In *Proc. AAAI*, 2021.
[12] X. Huang *et al. TabTransformer: Tabular Data Modeling Using Contextual Embeddings*. arXiv:2012.06678 (2020).
[13] E. Song *et al. AutoInt: Automatic Feature Interaction Learning via Self-Attentive Neural Networks*. In *Proc. CIKM*, 2019.
[14] T. Chen and C. Guestrin. *XGBoost: A Scalable Tree Boosting System*. In *Proc. KDD*, 2016.
[15] R. Agarwal et al. Neural Additive Models: Interpretable Machine Learning with Neural Nets. In Adv. Neural Inf. Process. Syst. 33, 2020.