https://doi.org/10.1038/s44387-025-00056-0

Escaping the forest: a sparse, interpretable, and foundational neural network alternative for tabular data

Salvatore Raieli[1,2,3] ✉, Nathalie Jeanray[1], Stéphane Gerart[1], Sebastien Vachenc[1] & Abdulrahman Altahhan[2,3] ✉

Tabular datasets are pervasive across biomedical research, powering applications from genomics to clinical prediction. Despite recent advances in neural architectures for tabular learning, there remains no consensus on models that balance performance, interpretability, and efficiency. Here, we introduce sTabNet, a meta-generative framework that automatically constructs sparse, interpretable neural architectures tailored to tabular data. The model integrates two key components. First, automated architecture generation leverages unsupervised, feature-centric Node2Vec random walks to define network connectivity, introducing a priori sparsity and improving generalisation while mitigating overfitting. Second, a dedicated attention layer jointly learns feature importance with model parameters during training, providing intrinsic interpretability. Evaluated across diverse biomedical tasks-including RNA-Seq classification, single-cell profiling, and survival prediction, sTabNet achieves performance on par with, or exceeding, leading tree-based models such as XGBoost, while remaining computationally efficient and CPU-trainable. Our experiments show that sTabNet generalises effectively across in-domain and out-of-domain datasets, yielding biologically consistent insights and surpassing post-hoc explainability methods such as SHAP in stability and clarity. Together, these results establish sTabNet as a foundational and versatile framework for data-efficient, interpretable neural learning on tabular data.

Although tabular data are among the most prevalent data types in scientific and industrial domains, recent advances in artificial intelligence have largely concentrated on images, text and multimodal tasks, with comparatively less emphasis on tabular learning[1,2]. Within biomedicine, however, machine learning has been extensively applied across diverse modalities, including gene expression, protein sequence post-translational modification prediction, disease risk modelling and mental health outcomes. In these areas, models such as gradient-boosted trees (e.g. Random Forest, XGBoost, LightGBM) have consistently delivered strong performance[3–6]. This has reinforced the position of tree-based methods as robust baselines for tabular problems, particularly in small- to medium-sized datasets where they frequently outperform conventional neural architectures[7].

At the same time, deep learning approaches for tabular data are increasingly being explored. A natural advantage of neural networks is their ability to transfer knowledge across tasks via fine-tuning, an attribute that has transformed progress in computer vision and natural language processing. In the tabular domain, two major directions have emerged. The first seeks to design tabular-native architectures that operate directly on raw feature spaces; while several large models have been proposed, they often incur high computational costs without consistently surpassing tree-based ensembles[7,8]. The second approach pursues tabular-to-image transformations, where feature similarity layouts (e.g. PCA, UMAP, t-SNE) enable the application of convolutional neural networks (CNNs) or ViTs to transformed tabular data. This strategy has demonstrated competitive or superior performance in specific domains, notably biomedicine, albeit at the expense of additional embedding steps. Together, these developments highlight both the promise and the limitations of current neural approaches, and underscore the need for simpler, efficient and high-performing neural network models that advance direct tabular learning while remaining complementary to representation-driven pipelines. In this work, we address this need by proposing a simple, efficient and high-performing neural network model tailored for direct learning from tabular data.

[1]Oncodesign Precision Medicine, Dijon, France. [2]School of Computer Science, University of Leeds, Leeds, UK. [3]These authors contributed equally: Salvatore Raieli, Abdulrahman Altahhan. ✉e-mail: salvatore.raieli2@gmail.com; a.altahhan@leeds.ac.uk

For neural models to represent a viable alternative in the tabular domain, they must not only perform competitively with tree-based ensembles but also provide interpretable outputs that are meaningful in biomedical settings. Interpretability is particularly important in domains such as genomics, proteomics and clinical prediction, where model decisions can influence biomarker discovery, risk assessment and therapeutic strategies[9,10]. In current practice, post-hoc attribution methods such as SHAP or Grad-CAM are widely used and have proven valuable, often producing intuitive visualisations. However, these are not the only options: attention-based mechanisms represent an alternative strategy, embedding feature weighting directly within the model architecture. While we acknowledge that attention should not be viewed as a definitive measure of feature importance, we employ it here as an architecture-level design that offers inherent transparency without the need for separate attribution steps.

At the same time, many existing neural approaches for tabular data rely on densely connected networks with large parameter counts. These architectures can model complex functions but also carry a heightened risk of overfitting, particularly in small- to medium-sized biomedical datasets[2,11,12]. Dense models can capture spurious correlations and memorise outliers or noise, resulting in poor generalisation. This motivates the exploration of alternative neural designs that integrate efficiency and a degree of interpretability at the architecture level, offering a different pathway than post-hoc explanations toward building more trustworthy tabular models.

Widely used convolutional architectures such as ResNet, MobileNet and VGG were designed for vision tasks, while large-scale foundational models like GPT and BERT target language and multimodal data. Neither class of models is well-suited for tabular problems: medium-sized CNNs rely on spatial priors absent in tabular feature spaces, and large foundational models carry parameter counts and training requirements far exceeding what is feasible for typical tabular datasets[13]. Although tabular-to-image pipelines can repurpose CNNs through feature similarity layouts, these approaches depend on additional embedding transformations and are therefore conceptually distinct from efforts to design tabular-native architectures.

Whether we are dealing with large or compact architectures, sparsity represents another promising direction, as highlighted by the lottery ticket hypothesis. The hypothesis suggests that many network weights can be pruned without compromising performance[14]. In practice, sparsity is usually induced post-training through pruning or related methods[15]. TabNetworks, for example, have leveraged sparsity guided by prior biological knowledge for genomics applications[16], yet such strategies remain domain-specific and difficult to generalise. Moreover, while sparsity may reduce model complexity, it does not address the broader challenge of interpretability. These limitations underscore the need for tabular-native neural architectures that jointly optimise efficiency, generalisability and interpretability.

Interpretability is especially critical in applications such as biology, banking and insurance, where algorithmic predictions directly affect human quality of life[17]. In biomedical settings, model transparency underpins biomarker discovery, drug target identification and risk factor analysis; in finance and insurance, it underlies fairness, compliance and accountability. Moreover, as governmental institutions consider regulations for AI-driven decision systems, the demand for interpretable tabular models is both a scientific and a societal imperative. This underscores the need for neural architectures that not only achieve competitive performance but also provide clarity in their decision-making processes.

In this paper, we introduce a foundational neural architecture designed to address key shortcomings of existing approaches for tabular data by combining efficiency, sparsity and built-in interpretability. This is particularly relevant for biomedical applications, where high dimensionality and limited sample sizes are common, and model transparency is essential for deriving actionable biological or clinical insight. Our model, sTabNet, enforces sparsity prior to training by defining feature-neuron connections through either domain knowledge or unsupervised feature graph exploration, thereby avoiding reliance on post-hoc pruning. It further incorporates an attention mechanism that yields feature weighting directly during training, supporting model transparency at the architectural level. Our contributions are threefold: (1) we introduce a principled method to impose sparsity using either external knowledge or graph-based random walks, (2) we develop a built-in attention mechanism to support tabular interpretability and (3) we demonstrate the model's transferability and performance on challenging biomedical tasks, including single-cell classification, multi-omics data fusion and survival analysis.

Results

In ref. 16, biologically constrained neural networks were introduced for biomedical applications as feed-forward neural networks (FFNNs) with sparsity imposed prior to training. Sparsity was achieved by using external biological databases to define the network architecture. Similarly,[18] employed the Reactome database to encode dependencies between genes by specifying their known interactions. In this framework, the data can be represented as a matrix of overlapping feature clusters (biological 'pathways', Fig. 1A). A binary adjacency matrix is then used to control feature interactions in the modified linear layer: entries are set to 1 if two features belong to the same pathway, and 0 otherwise. This binary matrix is applied element-wise to the FFNN weight matrix, constraining the network to respect known biological relationships (Fig. 1B).

The strength of this approach lies in its explicit integration of domain knowledge, which enhances model transparency[16]. However, post-hoc attribution methods such as SHAP are still required to identify important features, and the approach is inherently limited to domains with curated external knowledge bases. For datasets lacking such prior information, these biologically guided architectures cannot be straightforwardly applied.

Motivated by these limitations and inspired by the success of attention mechanisms, we propose a simple, tabular-native mechanism designed to be more amenable to interpretability. Our model, sTabNet, integrates sparsity before training and includes an attention module that provides direct feature weighting during learning. We show that sTabNet achieves competitive or superior performance compared to tree-based models on biomedical datasets, while offering clearer insights into relevant features than post-hoc approaches. Furthermore, by introducing a general algorithm, sTabAlgo, we extend these benefits beyond biomedical datasets, enabling the construction of sparse, interpretable tabular models in domains where external knowledge is unavailable. Remarkably, sTabNet, derived from sTabAlgo, remains competitive with tree-based models even without extensive hyperparameter tuning.

sTabNet: a sparse model for tabular data

We aim to develop a neural network architecture for tabular data that is high-performing, efficient and interpretable. Performance and efficiency are addressed through sparsity, while interpretability is achieved via an integrated attention mechanism, as described in the following sections.

To impose sparsity, we constrain the connectivity of the neurons in the architecture a priori, thereby reducing the number of trainable parameters and limiting the need for extensive tuning. This restriction helps prevent overfitting, lowers computational cost and enables different subsets of the network to specialise more effectively for different patterns in the data. Concretely, sparsity is introduced by grouping features according to shared innate properties and allowing them to connect only to neurons in subsequent layers that represent their group. This structured connectivity is illustrated in Fig. 1 and Supplementary Fig. 1. We can uncover the innate properties of features either by employing unsupervised learning methods, such as clustering or random walk, or by exploiting prior domain knowledge when it is available. In either case, we are grouping the features, not the data points. In other words, the intrinsic groups corresponding to neurons in the model are specified by the similarity of features, not the similarity of data points. A feature point consists of a vector; its components are the feature readings for all the data points in the dataset. Later, we present a comparison between using unsupervised learning and domain knowledge to construct the sTabNet model, demonstrating that, in general, models built using unsupervised learning perform comparably to those built using domain knowledge.

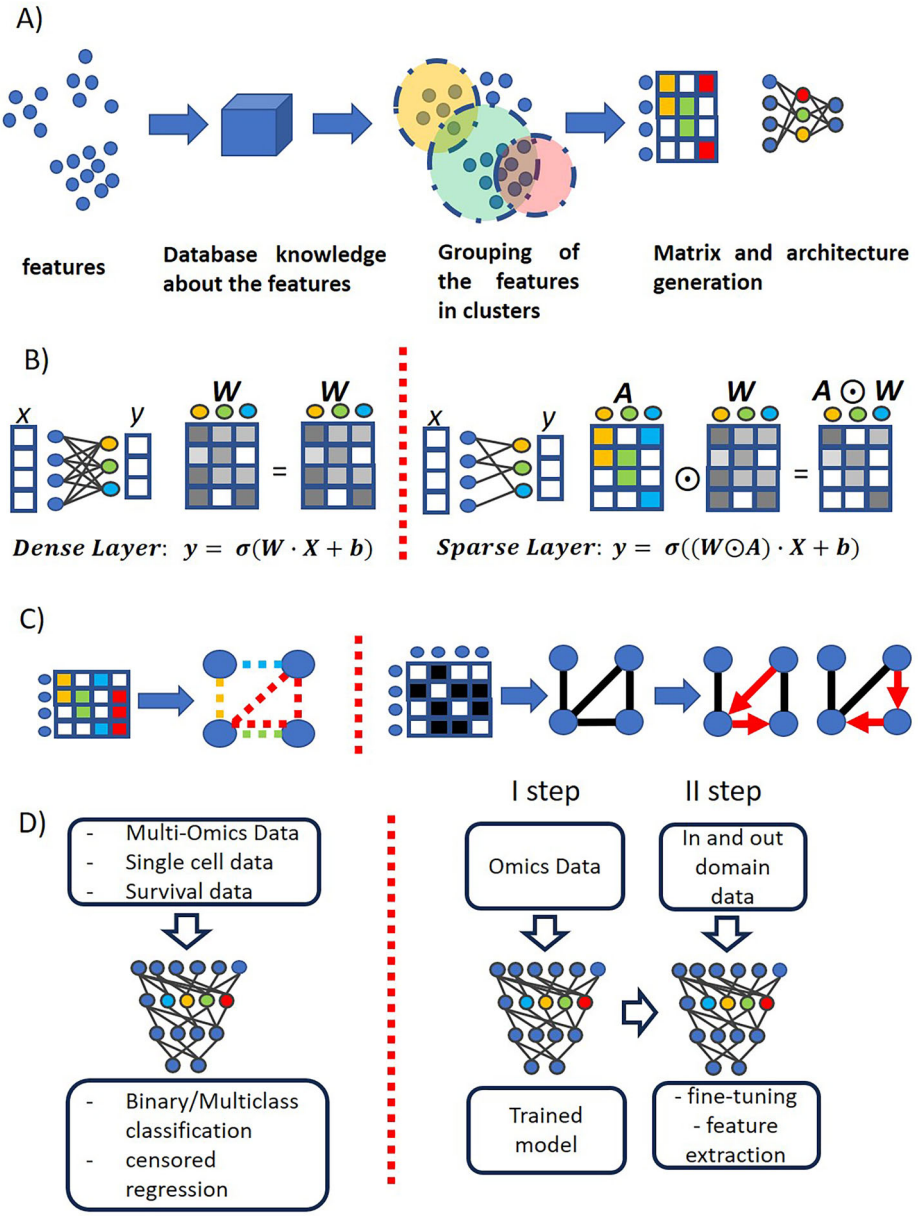


Fig. 1 | A sparse and interpretable neural network. **A** sTabNet Architecture: Features can be grouped according to prior knowledge or by using unsupervised learning (clustering or random walk) to build a matrix A where rows are features and columns are clusters (neurons). In this sparse model, a feature is connected to a neuron (which represents a cluster) only if it is a member of the cluster. **B** sTabNet Sparsity: The representation and the definition of the classical dense layer (left) and of the proposed sparse layer (right). The sparse layer is identical to the dense layer except for the Hadamard product between the weight matrix $\boldsymbol{W}$ and the matrix $\boldsymbol{A}$. **C** sTabNet Grouping: (left) The matrix $\boldsymbol{A}$ can be intended as a compressed view of an adjacency matrix of the feature graph. The neuron can also be defined as a random walk in the feature graph, thus learning a local approximation of the neighbourhood of a feature. Alternatively, one can use clustering (not shown in the figure). (right) Unrolling of the process on the left: When information about features in a dataset is not present, we calculate the cosine similarity matrix of the features. We assigned an edge between two features if their similarity is higher than 0.5. We performed random walks on the obtained graph and used the obtained random walks to build the sparse matrix in the modified layer. **D** sTAbNet as a Tabular Foundational Model: A scheme of sTabNet used for different tasks and data types. The same architecture can be used for common and challenging biological tasks (binary/multiclass classification, censored regression) and complex data (RNA-seq, single-cell and multi-omics data). sTabNet has been tested with real-world datasets for all these tasks. On the left, we are showing that the model can be trained on a dataset, and then the trained model can be used for other datasets and tasks through fine-tuning or feature extraction.

If $\boldsymbol{X} \in \mathbb{R}^{m \times n}$ is a tabular dataset, where m denotes the number of data points (rows) and n the number of features (columns), we define a design matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ that encodes the relationships among features. Our grouping algorithms operate along the feature dimension (n) to form a set of $K < n$ clusters, where each cluster corresponds to a group of related features. This contrasts with conventional clustering methods, which are typically applied along the sample dimension (m) to group data points rather than features. This grouping process dictates the number of neurons in the model, specifying the connectivities between the input layer and the first hidden layer. Features connected to the neurons are only those that belong to the group/cluster represented by this neuron.

When prior information about the features is available, it can be used to construct a binary matrix that controls the structure of the neural network (Fig. 1A). For instance, in RNA expression datasets, pathway databases can define which features are grouped together. This binary matrix is then applied as a mask to the weight matrix of the neural network, thereby enforcing sparsity (Fig. 1B).

In the absence of specific domain knowledge, the matrix can be derived through data-driven grouping. Considering feature similarity (e.g. cosine similarity), we either apply standard clustering algorithms or construct a feature graph and use random walks to generate the binary mask. This process enables the exploration of global or local feature interactions, depending on the chosen hyperparameters (Fig. 1C, D and Supplementary Fig. 1).

Supplementary Fig. 2 shows that this approach substantially reduces the parameter count. In Supplementary Fig. 3, we benchmark sTabNet against XGBoost across increasing feature dimensionalities (100–50,000) using synthetic datasets ($N = 6000$). sTabNet demonstrates superior scalability, particularly in terms of wall-clock time and peak training memory. Collectively, these results indicate that sTabNet achieves stable compute costs, fixed parameterisation and lower memory demands, making it well-suited to high-dimensional feature spaces. Finally, an ablation study (Supplementary Fig. 4) demonstrates that induced sparsity serves as a form of regularisation, and not performing conventional regularisation techniques, such as dropout, has only a minimal effect on performance.

sTabNet interpretability for tabular data

Next, we introduced a lightweight attention mechanism tailored for tabular data to derive feature importance directly during training, thereby eliminating the need for post-hoc attribution methods. In this formulation, the model learns feature relevance as part of the optimisation process, providing

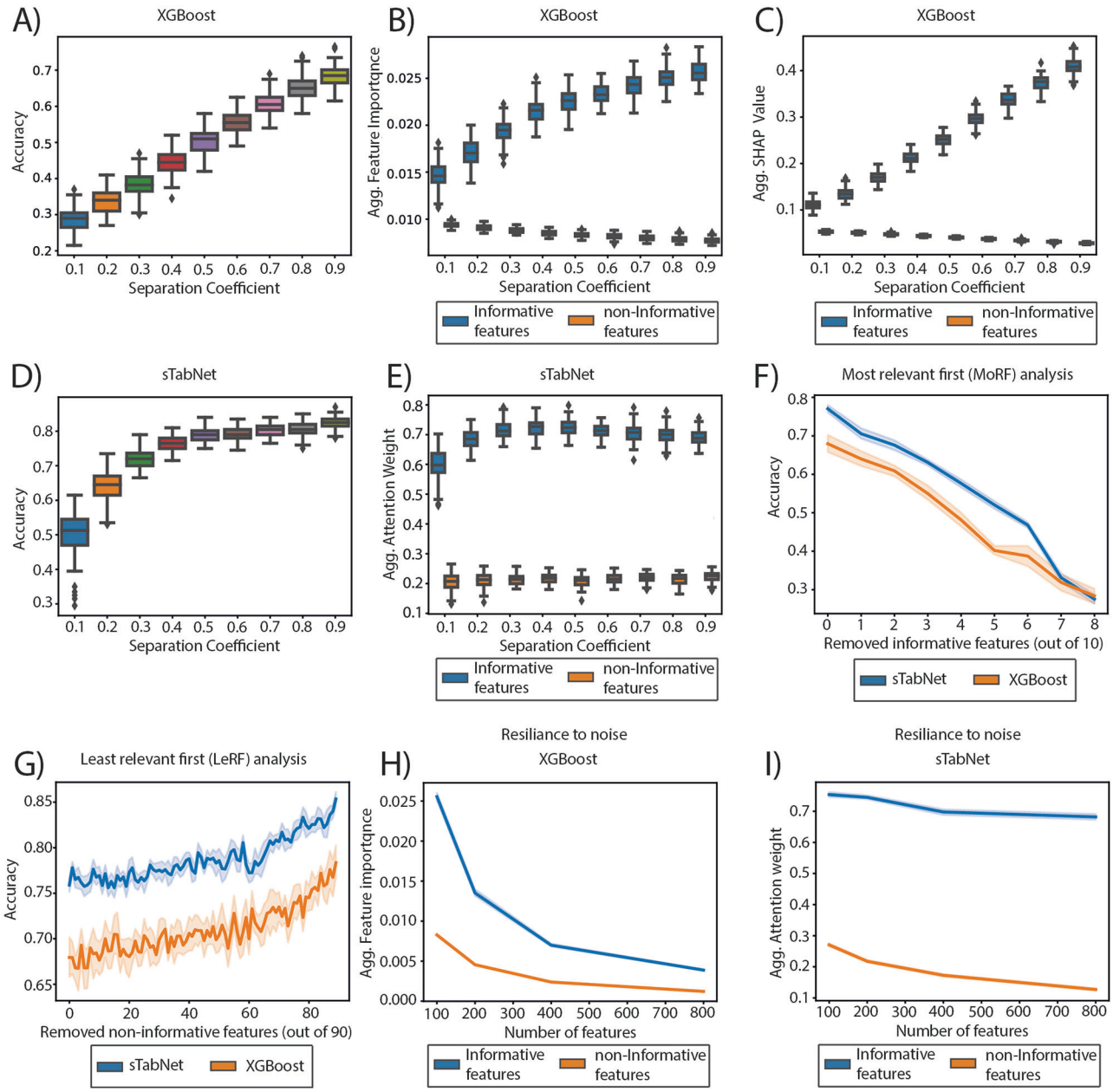


Fig. 2 | Attention mechanisms are a measure of feature importance. **A**–**F** Each boxplot represents a 100-fold hold-out validation; a lower coefficient represents a harder multi-classification task. **A** Multi-classification accuracy in XGBoost with an increase in separation difficulty. **B** Separation between the average importance weight (XGBoost's feature importance) assigned to real informative and non-informative features. **C** Separation between the average importance weight (SHAP value) assigned to informative and non-informative features. **D** Multi-classification accuracy in sTabNet with an increase in separation difficulty. **E** Separation between the average importance weight (feature attention weight) assigned to real informative and non-informative features. **F**, **G** Accuracy is plotted for each different model (XGBoost and sTabNet). The shade represents the standard deviation (10 different models for each removed feature). **F** MoRF analysis. **G** LeRF analysis. **H**, **I** The shade represents the standard deviation (10 different models trained). **H** Feature importance for XGBoost when increasing the number of non-informative features. **I** Feature importance for the sTabNet with an increase in the number of non-informative features.

transparency without additional interpretability steps. We evaluated sTabNet to confirm that the attention mechanism accurately captures the contribution of individual features across both simulated and real datasets. Furthermore, we assessed sTabNet on a range of complex biomedical tasks —including multi-omics integration, single-cell analysis, multiclass classification and survival regression—and compared its performance with that of tree-based models (Fig. 1D).

Interpretability, attention and feature importance

As highlighted by ref. 1, interpretability measurements are complicated because we do not have a dataset where the importance of the features is known in advance (a ground truth). Therefore, to test the effectiveness of our interpretability approach, we used a set of synthetic datasets with various difficulties. In this evaluation framework, the ground truth of feature importance and the level of complexity can be controlled a priori via redundancy and separation coefficients, respectively. Figure 2 shows this set of experiments.

First, from a classification perspective, we observed an expected linear decrease in XGBoost accuracy with the increase in the difficulty of the multi-classification task (Fig. 2A), where the classes are harder to separate. When the classes are poorly separated (separation coefficient 0.1), we observed a reduction in the weight assigned to important features, making it harder to

separate them from noisy features (Fig. 2B), and the models were almost performing random guessing. The SHAP value follows the same pattern for the feature importance, not allowing better identification of important features (Fig. 2C). Thus, SHAP explanations in a sense are the expression of the prediction abilities of XGBoost, they are dependent on its accuracy, but they do not improve the ability of the model in capturing the feature importance (and to separate informative and noisy features). We incorporate a tabular attention mechanism (defined in the Methods section) to automatically determine the importance of the input features. While we also observed a decrease in sTabNet performance with increasing difficulty in the multi-classification task, the reduction is less dramatic (Fig. 2D and Supplementary Fig. 5A–C). Moreover, the separation between important features and noisy features is more pronounced (Fig. 2E). Interestingly, when the classification task is more complex, sTabNet needs more training epochs to reach better performance and better separation (Supplementary Fig. 6A, B).

We conducted a feature ablation study to evaluate the relationship between attention weights and feature importance. Our hypothesis is that attention weights can serve as proxies for the true importance of individual features. The Most Relevant First (MoRF) analysis revealed a sharp performance decrease when features with high attention weights were removed, indicating that the attention mechanism effectively identifies the most discriminative features (Fig. 2F). While acknowledging that attention should not be regarded as a definitive explanation of model behaviour, these findings suggest that it offers a quantitative and empirically verifiable measure of feature relevance.

The Least Relevant First (LeRF) analysis indicates that attention scores effectively identify noisy features (Fig. 2G). As the number of noisy features increases, the corresponding feature importance estimated by XGBoost declines markedly (Fig. 2H). In contrast, the attention-based feature importance in our proposed sTabNet model remains stable, enabling a clearer distinction between informative and noisy features (Fig. 2I).

Finally, we compared our approach with other commonly used feature importance methods and observed a general concordance (Supplementary Table 1). These results indicate that the attention weight can serve as a reliable feature importance score and acts as a stronger discriminator between informative and noisy features than both XGBoost feature importance and SHAP values (Fig. 2C). Moreover, the attention-based scores exhibit greater stability in distinguishing significant from noisy features across different datasets.

sTabNet on complex real-world datasets

We evaluated our model on complex multi-omics datasets, which pose unique challenges for machine learning models[19–22]. Our sTabNet architecture was constructed using the METABRIC multi-omics dataset, as described earlier. Because there is no consensus in the literature regarding the most suitable activation function for biologically constrained networks[18,23], we systematically tested several activation functions during training. However, no significant differences in performance were observed across activation choices (Supplementary Fig. 7A).

More importantly, we observed that a standard FFNN trained on this dataset predominantly predicted the majority class, indicating a severe sensitivity to class imbalance. Consequently, we did not include further experiments with fully connected architectures in this study. Similarly, CNNs did not achieve satisfactory performance on this dataset (Fig. 3A).

sTabNet yields results comparable to those of XGBoost (Fig. 3A). Furthermore, prepending an attention layer results in a slight performance increase (Supplementary Fig. 7B and Supplementary Table 2). Since we do not have the ground truth for feature importance for this dataset, we have selected the 100 features with the highest attention weights, and we conducted a gene-set enrichment[24] for diseases on this feature subset. The results show that the features with the highest attention weights are attributed to cancer-related genes (Fig. 3E, F and supplementary Table 3), which is consistent with the results from the synthetic dataset.

sTabNet generalisation and transfer learning

Since one advantage of neural networks is their ability to be fine-tuned for new tasks, we investigated whether a model trained on METABRIC (multiclass classification) could be adapted to a binary classification task using the TCGA-BRCA dataset. The fine-tuned model successfully adapted to both a different objective and an in-domain dataset, demonstrating effective in-domain transfer (Fig. 3B). Moreover, a frozen sTabNet model extracted meaningful representations that could be used to train a linear classifier, further supporting its representational capacity (Fig. 3B). We also observed promising out-of-domain adaptation on TCGA-LUAD, where both fine-tuning and feature extraction yielded good performance (Fig. 3B).

Learning-curve analyses revealed that sTabNet maintains high performance even with limited training data. Classification accuracy and precision-recall metrics remained stable across varying dataset fractions, achieving over 90 % AUROC and AUPRC using only 10 % of the available data (Supplementary Fig. 8). These results indicate that sTabNet is highly data-efficient, achieving strong generalisation from small samples with limited variability across random seeds.

We further evaluated cross-dataset generalisation under a Leave-One-Dataset-Out (LODO) protocol using five independent gene expression datasets (Lung GSE19804, Lung GSE18842, Leukaemia GSE63270, Throat GSE42743). Models were trained on each dataset and evaluated both on the held-out test partition and directly on the remaining datasets (Supplementary Fig. 9A, B). The aggregated results indicate that sTabNet learns transferable representations across related biological domains. Transfers between similar tissues (e.g. lung-to-lung) yielded the highest performance, while transfers between biologically dissimilar tissues (e.g. solid-to-blood cancers) resulted in lower accuracy, reflecting underlying biological heterogeneity. In the pooled-source LODO setting, performance remained robust for lung datasets, moderate for throat and lower for leukaemia (Supplementary Fig. 10), suggesting that transfer success correlates with tumour-type similarity.

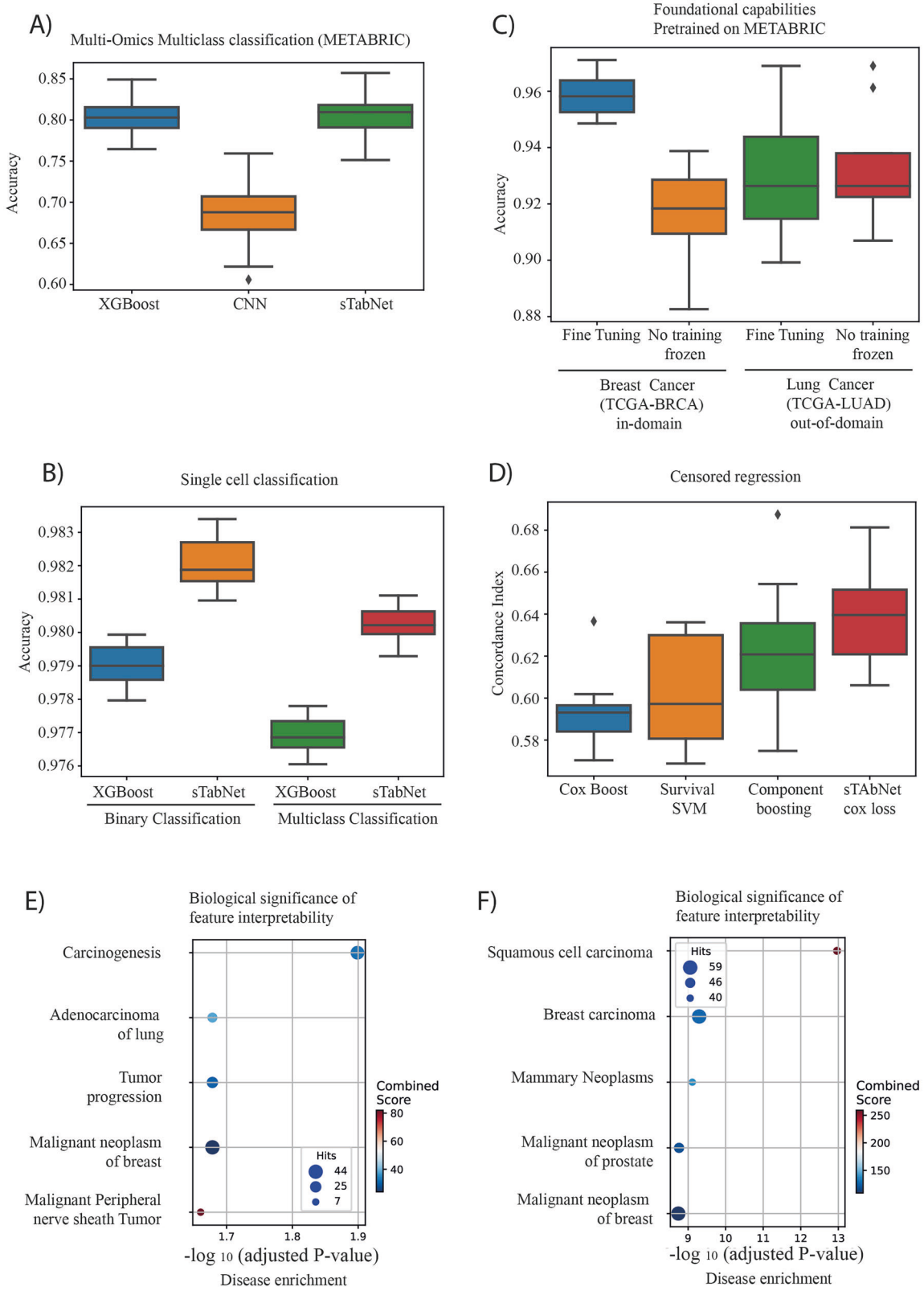
Comparison with previous tabular models

Gene expression datasets remain a major challenge for machine learning models because they typically contain thousands of features but only a few dozen samples. Such datasets represent an extreme case of the tabular domain, characterised by high dimensionality, noisy and unbalanced features, and complex decision boundaries. Several neural architectures, including TabNet and NODE[25–27], have demonstrated competitive performance on tabular benchmarks. Other strategies have focused on feature selection or neuroevolution methods to address the extreme feature-to-sample ratio[28–33]. A complementary line of research transforms tabular data into images, enabling the application of convolutional models such as DeepInsight and its derivatives[34–36].

Our approach differs from these models in three key aspects: (1) it enables the integration of domain knowledge-such as curated biological pathways-directly into the network structure; (2) it introduces sparsity a priori in the architecture, rather than learning it post-hoc or through evolutionary search; and (3) it provides native interpretability via built-in attention mechanisms, eliminating the need for separate explanation models.

We benchmarked sTabNet against seven classification approaches (XGBoost, TabNet, NODE, DeepInsight-PCA, DeepInsight-tSNE and others) using a subset of the Cumida benchmark expression dataset. sTabNet achieved statistically superior performance across ten independent expression datasets (Supplementary Fig. 11). Furthermore, as reflected by F1 scores, models such as TabNet and NODE struggled with class imbalance-a common characteristic of expression datasets-whereas sTabNet maintained consistently higher performance.

To evaluate interpretability and biological relevance, we compared sTabNet and DeepInsight feature attributions (Supplementary Fig. 12). Gene relevance was quantified using two complementary measures: the number of PubMed articles mentioning each gene in the context of 'liver cancer' or 'HCC', and scores assigned by a large language model (LLM) assessing gene relevance based on expression statistics, Gene Ontology (GO) pathways and representative abstracts. Both the PubMed and LLM analyses indicate that



sTabNet attention scores and integrated gradients (IG) align more closely with known biological evidence than DeepInsight-CAM (Supplementary Fig. 13). sTabNet achieved the highest median LLM relevance score and the largest proportion of genes scoring ≥4, supporting the conclusion that it yields more biologically plausible gene rankings than image-based approaches while maintaining strong predictive accuracy (Supplementary Fig. 14).

sTabNet on single-cell technology

Single-cell technologies have profoundly advanced our understanding of development, cell identity and disease mechanisms. However, single-cell datasets present unique computational challenges, typically featuring far more variables than samples and lacking consensus strategies for effective modelling[37–39]. To evaluate sTabNet in this setting, we used single-cell RNA-

Fig. 3 | sTabNet provides a foundational model to perform in-domain and out-of-domain fine-tuning, it is interpretable and outperforms tree-based models in general. A Comparative table between XGBoost, sTabNet and CNN on the METABRIC multi-omics dataset (multiclass classification). Performance comparison across models. Statistical significance between model performances was assessed using pairwise Wilcoxon signed-rank tests. Reported p values comparison: sTabNet with XGBoost ($p = 0.035$) and with CNN model ($p < 1e\text{-}84$). **B** in-domain (breast cancer) and out-of-domain (lung cancer) adaptation. The model was trained on the Metabric dataset, then fine-tuned on other datasets. Accuracy on TCGA-BRCA (same domain as the METABRIC dataset) and TCGA-LUAD (different domain from the original dataset) for fine-tuning or feature extraction. **C** Binary and multiclass classification accuracy for sTabNet and XGBoost on single-cell data (breast cancer (GSE161529)). **D** Concordance index for METABRIC survival analysis. Statistical significance between model performances was assessed using pairwise Wilcoxon signed-rank tests. Reported p values comparison: sTabNet with Component boosting ($p = 0.029$), with Survival SVM ($p = 0.011$), with Cox Boost ($p < 0.01$) **E** Disease enrichment for the 100 top genes according to attention importance. **F** Disease enrichment for the 100 top genes according to attention importance from METABRIC.

seq data from the Tumour Immune Single-cell Hub 2 (TISCH2; GSE161529) for breast cancer. sTabNet was applied to both binary classification (tumour versus normal cells) and multiclass classification (cell type prediction). Ten-fold cross-validation was performed for each model, and average accuracies were reported.

After freezing the trained model, we extracted the learned representations and visualised them using UMAP. sTabNet demonstrated strong performance across both binary and multiclass tasks (Fig. 3D and Supplementary Table 4). Moreover, the learned embeddings showed clear separation of cell populations (Supplementary Fig. 15), consistent with the hypothesis that reduced connectivity enables different subnetworks to specialise in distinct data patterns. These results indicate that sTabNet captures biologically meaningful features that can be leveraged for downstream analyses. Importantly, given the high dimensionality of genomic data[40], sTabNet can also serve as an effective feature selection tool, reducing both feature count and model complexity within analysis pipelines (Supplementary Fig. 16).

sTabNet and survival analysis

Survival analysis is a branch of machine learning concerned with modelling the relationship between the time to an event (such as death or system failure) and predictive features[41,42]. Although it can be framed as a subcase of a regression problem, traditional regression methods often fail to capture the underlying temporal and nonlinear dependencies[43]. Given the complexity and biological importance of survival data, several models have been developed for this task, such as the Cox proportional hazards model[44,45]. However, these approaches struggle to model nonlinear effects and to scale effectively with large datasets.

To address these challenges, we compared sTabNet with state-of-the-art methods from the Scikit-survival library, which includes several ensemble and kernel-based algorithms[46]. sTabNet consistently outperformed these baselines, including ensemble tree-based and support vector machine-based models (Fig. 3D). In addition, the attention mechanism in sTabNet enabled identification of the most influential predictors of survival (Supplementary Fig. 17).

Overall, these results demonstrate that sTabNet with integrated attention mechanisms not only surpasses tree-based models on genomic survival datasets but also offers interpretability through feature-level attribution. By learning meaningful latent representations, sTabNet supports both in-domain and out-of-domain adaptation in a unified framework. Moreover, its versatility extends to a range of complex biological tasks, including multi-omics classification, single-cell analysis and survival regression, as illustrated in Fig. 3.

Unsupervised sTabNet and expression data

While the use of biological knowledge (e.g. Gene Ontology) to define network connectivity enhances interpretability and biological plausibility, such databases are manually curated and may be incomplete, outdated, or biased toward well-studied genes. Moreover, not all genes are annotated, and inclusion criteria vary across resources, limiting both coverage and generalisability. To address these limitations, we also evaluated an unsupervised alternative that automatically determines the connectivity of sTabNet based on feature-similarity graphs combined with random walks, thereby removing dependency on prior knowledge and enabling application to any tabular dataset. When applied to complex genomic data, the random-walk-based approach yielded performance comparable to the Gene Ontology-based configuration, demonstrating its scalability and broad applicability (Fig. 3F, Supplementary Fig. 18 and Supplementary Table 6).

sTabNet vs. tree-based models for tabular data

Although different transformer-based models have been tested, different researchers suggest that the high capacity of neural networks hinders their applicability to tabular tasks[47]. Since information about the features is often not available, we described a method to leverage sTabNet to address problems in arbitrary tabular datasets. In fact, the aim was to identify a simple method to impose sparsity before training while being competitive with tree-based models (Fig. 4A). Despite not conducting a hyperparameter search and using a simple architecture, the sTabNet is shown to outperform the tree-based model: median accuracy 0.71 versus 0.70 (Fig. 4B and supplementary Table 5 and Supplementary Figs. 19 and 20).

Within a sparse neural network layer, each neuron's connections are determined by a random walk across a feature graph. This process ensures that the neuron establishes connections only within a localised, harmoniously related subset of features (i.e. its neighbourhood) on the graph. We noticed that the top random walk (selected using an additional attention layer) always represents the same neighbourhood (Fig. 4C, D). We conducted an ablation study on the features by removing the top 5 features to verify and study the effect of these features (in connection with their neurons). Figure 4E shows that removing the involved feature impacts accuracy. Interestingly, removing these features increases the number of false positives compared to random feature removal (Fig. 4F). We did not observe an increase in false negatives with feature ablation (Fig. 4G). These results suggest that local patterns in the feature graphs can be associated with a particular class, thereby contributing to the emergence of modularity. The recall ability of a model is dependent on the interconnectivity of its features collectively, while the precision is dependent on the existence or absence of a certain set of features in a pattern. Our results consistently show that the model is able to identify these features that are important for the ability to precisely classify a pattern, which contributes to better precision.

Discussion

Tabular data are among the most prevalent data types across scientific and industrial domains, underpinning applications in medicine, psychology, finance, cybersecurity and user modelling[1,48–52]. While recent years have seen growing interest in applying neural networks to tabular problems—including transformer-based and tabular-to-image approaches—the field still lacks consensus on architectures that balance performance, interpretability and computational efficiency[1,53]. In this work, we argue that the challenge lies not in the absence of neural approaches but in their tendency toward overparameterisation and suboptimal scaling for small or medium-sized datasets. To address this, we introduced a simple yet effective architecture that leverages intrinsic feature relationships—or domain knowledge when available—to impose sparsity a priori within the network structure, thereby reducing model complexity while preserving both accuracy and interpretability. In this work, we employed attention mechanisms to quantify feature importance within neural networks. The role of attention as an explanatory tool remains a subject of active debate, with several studies questioning whether attention weights reliably reflect model reasoning[54–56]. Aware of these concerns, we designed a specialised attention mechanism tailored to tabular data and systematically evaluated it using both synthetic and real-world datasets.

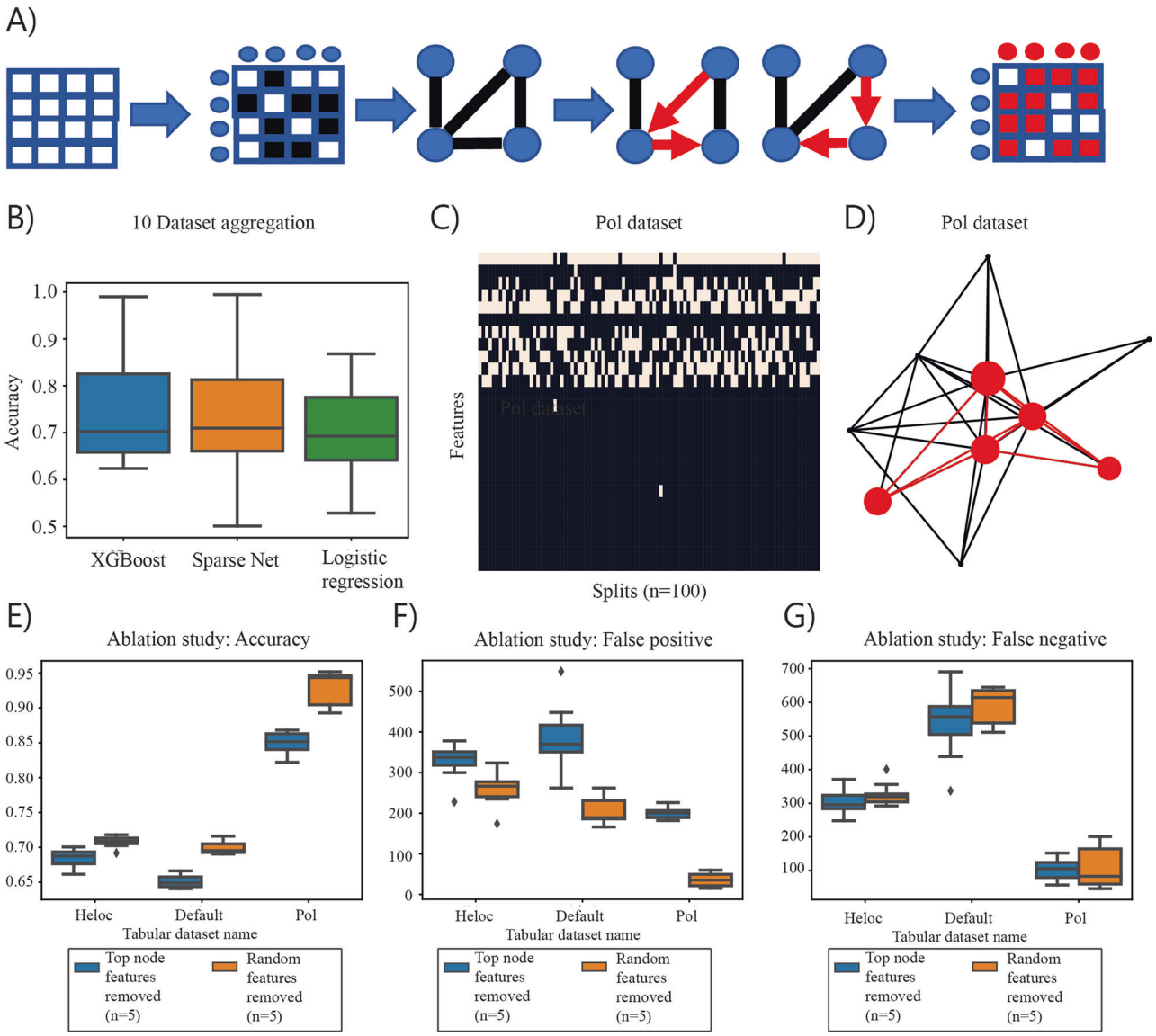


Fig. 4 | Generality of the proposed sTabNet architecture. sTabNets are competitive with tree-based models for tabular data. **A** The algorithm that involves a random walk process is used to adapt the model to any tabular dataset. We calculated the similarity between the dataset's features (cosine similarity matrix) and used the obtained matrix to generate a feature graph. We conducted random walks on the feature graphs to explore the local neighbourhood of each feature. We use that knowledge to build the sparse matrix of the neural network **B** Binary classification accuracy for the tabular benchmark (100 models for 3 techniques: we generated 10 models corresponding to 10 runs -each with a random training/testing split- for 10 datasets). **C** Feature presence in the top random walks (100 experiments on the pol dataset with a random training/testing split) **D** Feature graph for the pol dataset (isolated nodes are removed), the red-highlighted nodes represent the top 5 features present in the top random walk. Ablation study: accuracy (**E** plot) or false positives (**F** plot) or false negatives (**G** plot) performance for three datasets, removing five random features vs. removing the 5 top features of the random walk process (10 experiments for each dataset).

Using a synthetic dataset with predefined feature importance as ground truth, we demonstrated that attention-derived importance scores effectively distinguish informative features from noise. In patient datasets, the top-ranked features identified by attention values corresponded to biologically and clinically relevant variables, further supporting the view that attention can serve as a meaningful explanatory signal. When applied to real-world datasets with domain knowledge (such as cancer datasets), this approach enables the identification of key features and facilitates exploratory insights into the data. This capability is broadly valuable across domains-from biomedicine to finance-where understanding which features drive predictions is essential. Furthermore, we showed that sTabNet can be effectively constructed even in the absence of prior knowledge, and that this data-driven variant consistently produces reliable and competitive results across multiple benchmarks.

In addition, we showed that it is essential to evaluate interpretability using datasets where the ground truth of feature importance *is known*. Our findings highlight that even established attribution methods may be unreliable, particularly for complex tasks involving a large number of features. Furthermore, attention weights can be leveraged for feature selection to remove non-informative variables, improving both model efficiency and robustness. This capability is especially valuable for real-world datasets, which are often noisy and contain redundant features.

Most previous comparisons between tree-based models and neural networks have focused on relatively simple datasets with limited numbers of features[7]. Such datasets are not representative of real-world conditions, particularly in biologically complex domains. In contrast, we demonstrated that when modelling tasks require learning rich, high-dimensional data representations, sTabNet performs competitively with tree-based models.

We selected biological datasets because they exemplify real-world data characterised by intrinsic complexity, nonlinearity and noise, demanding models capable of adaptive representation learning[57]. These datasets also contain a large number of irrelevant or redundant features, posing significant challenges for feature selection and model robustness[58].

Moreover, in such complex tasks, neural networks offer unique advantages: they can learn transferable representations, enable end-to-end optimisation and integrate feature extraction and selection within a unified framework. sTabNet embodies these capabilities, demonstrating foundational adaptability across heterogeneous tabular problems. While our experiments focused on biomedical data, the same principles extend naturally to other domains—such as psychology and customer analytics—where structured yet complex data are prevalent.

While several sparse models have been proposed for cases where external domain knowledge is available, we introduced a generalisable approach that enables the construction of sTabNet for any tabular dataset. Our results show that sTabNet achieves superior performance compared to tree-based models. As previously suggested in the literature, dense architectures often exhibit excessive capacity for small or low-dimensional datasets, leading to overfitting. In contrast, systematically constraining the network through sparsity—imposed in an unsupervised and data-driven manner—proves effective in guiding the model toward learning meaningful patterns. Furthermore, this sparse formulation encourages neurons to specialise in local neighbourhoods within the feature graph, enhancing both interpretability and robustness. Interpretability remains a critical factor for the practical adoption of neural network models, particularly in domains where transparency and trust are essential.

It is worth noting that, in recent years, a wide range of models have been developed for analysing gene expression data. Many of these approaches rely on complex processing pipelines—such as transforming tabular data into image representations—or on advanced architectures including pre-trained CNNs, Transformers and Vision Transformers. While effective, these methods often require substantial computational resources and specialised infrastructure, such as high-performance GPUs[59–62]. In contrast, this work focuses on a compact neural network that can be efficiently trained on CPUs, offering computational efficiency within a single, end-to-end framework.

Importantly, our motivation extends beyond architectural efficiency to address the core challenges of biomedical data science-namely, high dimensionality, limited sample sizes and the imperative for interpretability. In genomics and clinical contexts, identifying the features that drive predictions is not merely desirable but essential for translational insight. sTabNet was explicitly designed with these constraints in mind: it enables the integration of prior knowledge when available, remains robust in purely data-driven settings and provides native interpretability through attention weights aligned with biologically meaningful features. This alignment between model design and domain needs distinguishes sTabNet from generic sparse or attention-based architectures not tailored to biomedical challenges.

While classical unsupervised approaches such as K-means clustering or dimensionality reduction techniques (e.g. PCA, t-SNE) can reveal patterns in tabular data, they are not designed to capture overlapping or localised feature interactions. In contrast, random walks on a feature graph facilitate a soft, overlapping grouping of features, whereby each neuron (random walk) samples a local neighbourhood of related variables. This approach mirrors the modular and redundant organisation commonly observed in biological systems, where genes and pathways often participate in multiple functional contexts. Moreover, unlike hard clustering methods, random walks introduce controlled stochasticity, enabling natural ensembling and the emergence of diverse architectural patterns. Such flexibility is particularly valuable when prior biological knowledge is unavailable.

However, our method also generates isolated nodes; in the future, other alternatives to developing the feature graph could be explored. For example, alternative distance measurements or a weighted Node2vec random walk can be explored. In addition, L1 regularisation (or other regularisation techniques) can be used to increase the network's sparsity. In addition, we conducted the benchmark on only a limited number of datasets and tasks, and in the future, we plan to add a larger number of datasets (RNA-seq, single-cell, multi-omics). In this direction, we would like to train STabNet with a large number of different datasets and test its capabilities on unseen diseases, tissue types, or noisy real-world data. Second, while the GO-based masks provide structural interpretability, we did not perform in-depth biological analysis of the learned pathway activations. In addition, future studies should also consider the translationality of the features identified in clinical settings and biological experiments.

Methods

Overall construction of the sTabNet approach

Our model, **sTabNet**, is generated via a meta-generative framework that constructs the architecture based on feature relationships. Two modes are supported: Knowledge-driven: If domain knowledge (e.g. Gene Ontology pathways or other sources) is available, a binary matrix $\boldsymbol{A}$ encodes known feature-group associations (i.e. connections between input features and hidden layer neurons). Unsupervised mode: If no such prior is available, we apply unsupervised learning (e.g. cosine similarity and Node2vec random walks) to build a feature graph, and sparsity is derived from local neighbourhoods (i.e. learned clusters of features). Each hidden neuron corresponds to a feature group or graph neighbourhood, and connections are created only within these defined subsets, thereby generating a sparse and interpretable connectivity map.

Sparsity is achieved a priori, before training, by defining a binary masking matrix $\boldsymbol{A}$ (also interpreted as a feature graph adjacency matrix). This mask is used in a Hadamard product with the neural network weight matrix:

$$\boldsymbol{H}_1 = \sigma(\boldsymbol{X} \cdot (\boldsymbol{A} \odot \boldsymbol{W}))$$

This ensures that only feature connections defined in $\boldsymbol{A}$ contribute to training. Compared to post-hoc pruning, this has several benefits: It reduces the number of learnable parameters upfront. It mitigates overfitting in high-dimensional settings. It allows the model to focus on local feature interactions, improving interpretability and robustness. We emphasise that sparsity is enforced on the feature dimension, not the data samples, distinguishing our approach from many compression techniques. We also explored L1 regularisation for further sparsification, confirming its utility for feature selection without degrading performance.

We designed this architecture to differ fundamentally from traditional dense networks and models that rely on post-training pruning. First, the architecture itself is generated automatically, either from domain knowledge or through unsupervised feature similarity, enabling a principled and data-driven definition of connectivity. Second, sparsity is imposed a priori, improving generalisation, reducing overfitting and enhancing interpretability. Third, a dedicated attention layer is incorporated to learn a measure of features importance jointly with model parameters, i.e during training, offering intrinsic interpretability. Finally, each neuron learns a localised feature interaction-analogous to modular graph representations-naturally aligning the model with the structured relationships inherent in biological data.

In contrast to architectures that learn sparsity through attention alone or rely heavily on post-hoc interpretability methods (e.g. SHAP), sTabNet integrates attention directly into its architecture, providing feature-level interpretability during training while maintaining computational efficiency.

Tabular attention mechanism

To enable feature-level interpretability within our sparse architecture, we introduce an attention mechanism adapted for tabular data. Unlike sequence or image attention, which models relationships across temporal or spatial dimensions, tabular attention focuses on identifying the relative importance of features for each data sample during training.

Let $\boldsymbol{X} \in \mathbb{R}^{b \times n}$ denote a batch of data points from a tabular dataset, where n is the number of features and b is the batch size. The attention layer learns an attention weight vector $\tilde{\boldsymbol{w}} \in \mathbb{R}^n$ that modulates the contribution of each feature within a data point. To enrich this representation and capture the holistic importance of a data point given its feature composition, we also

define an attention score matrix $\tilde{\boldsymbol{W}} \in \mathbb{R}^{n\times n}$ constructed additively via two projection vectors, $\tilde{\boldsymbol{w}}_1 \in \mathbb{R}^{n\times 1}$ and $\tilde{\boldsymbol{w}}_2 \in \mathbb{R}^{n}$, to operate across the feature space:

$$\tilde{\boldsymbol{\alpha}} = \tilde{\boldsymbol{w}} * \mathrm{softmax}\big(\varphi(\boldsymbol{X}, \tilde{\boldsymbol{W}})\big),$$

where all operations act along the feature dimension n, leaving the batch dimension b unchanged, and * denotes broadcasting. The function φ corresponds to a choice of attention scoring function[63–66]. In this work, we evaluated four variants: (i) a Bahdanau-inspired form, $\tanh(\boldsymbol{X}\tilde{\boldsymbol{W}})$[63]; (ii) the dot product, $\boldsymbol{X}\tilde{\boldsymbol{W}}$[64]; (iii) a cosine-similarity score, $\frac{\boldsymbol{X}\tilde{\boldsymbol{W}}}{\|\boldsymbol{X}\|\,\|\boldsymbol{W}\|}$[65]; and (iv) the scaled dot product, $\frac{\boldsymbol{X}\tilde{\boldsymbol{W}}}{\sqrt{n}}$[66], where n is the feature dimensionality.

The resulting score matrix $\varphi(\boldsymbol{X}, \tilde{\boldsymbol{W}}) \in \mathbb{R}^{b\times n}$ contains unnormalised relevance estimates for each feature within every data point in the batch. A row-wise softmax is then applied across the feature dimension to obtain normalised attention coefficients, $\tilde{\boldsymbol{\alpha}}$, which reflect the relative importance of each feature given the feature composition of that specific data point. The attention weights are subsequently used to rescale the input features, yielding an attention-weighted representation:

$$\tilde{\boldsymbol{X}} = \boldsymbol{X} \odot \tilde{\boldsymbol{\alpha}},$$

where $\odot$ denotes element-wise (Hadamard) multiplication. This operation amplifies the contribution of informative features while attenuating noisy or redundant ones. Importantly, because the attention weights are optimised jointly with the network parameters, feature importance is inferred *during training* rather than through post-hoc analysis.

Overall, the attention layer preserves the input dimensionality, allowing seamless integration into fully connected or sparse architectures. When combined with the sparse connectivity imposed by the binary mask $\boldsymbol{A}$, this mechanism enables the model to focus on meaningful local feature interactions while providing a direct and interpretable measure of feature relevance for each prediction.

Analysis of the interpretability of the sTabNetworks

To test whether the attention mechanism score can capture the importance of a feature in the dataset, we built simulated data for a multi-classification task. The multiclass dataset is built as in[67] and using the scikit-learn package implementation, where the difficulty of the classification task is regulated by a hyperparameter (separation coefficient, range 0.1–1). Where 0.1 is the most complex task for the model, and 1 is a trivial separation task. We defined informative and non-informative features (random noise) in the dataset. We used scikit-learn standard implementation (*make classification*), resulting in a matrix $\boldsymbol{X} \in \mathbb{R}^{b\times n}$ (where n is the sum of informative and non-informative features). The sTabNet and XGBoost have been then trained to optimise a multi-classification objective (6 classes), i.e. categorical cross-entropy as a loss function.

For each separation coefficient, we trained ten different models (1000 examples with 100 features, of which 10 informative and 90 non-informative) and measured the classification accuracy on the test set (0.2 of the dataset). We then examined whether a model can discriminate and separate informative features from random noise. For this purpose, we analysed the importance score (for XGBoost) or the attention weight assigned to each feature in the sTabNet. For XGBoost, we also used the SHAP value to determine the features' importance[68]. For the sTabNet, we analysed whether the attention weight can be considered a feature importance measure. As defined in ref. 69, we conducted LeRF and MoRF studies to assess the fidelity of the feature importance attribution. For XGBoost and SHAP, we used the standard parameters as defined in the XGBoost package. For the sTabNet, we used a simple architecture in Keras. The architecture has an input layer, an attention mechanism, a sparse layer with 100 neurons, a linear layer with arbitrary 64 neurons (to test the effect of attention only, not the effect of domain knowledge or unsupervised learning), and an output layer with softmax activation. We used dropout regularisation, ADAM optimiser and categorical cross-entropy as a loss. Since previous works focused on extensive hyperparameter optimisation[47] for the tabular neural networks, we intentionally did not conduct it to show performance with simple settings. For each separation coefficient, we randomly split 100 times (100 times cross-validation) and calculated the multiclass accuracy (both for XGBoost and the sTabNet). We extracted the feature importance attributed to each feature from the model.

For LeRF and MoRF analysis, we iteratively removed the most or least important feature according to the mean importance (mean of 10 models) and retrained on the reduced feature set. We also checked the stability for XGBoost and sTabNet by progressively increasing the number of noisy features (while keeping the number of important features constant to 10) and analysing the importance attributed to the features (informative and non-informative).

Application to multi-omics dataset, fine-tuning and feature extraction

Multi-omics data are notoriously complex to handle: they are expensive to collect, hetero-modals, and suffer the curse of dimensionality (for a dataset $\boldsymbol{X} \in \mathbb{R}^{m\times n}$, we have $n \gg m$)[19]. While designing an algorithm for this data is challenging, identifying important features leads to direct important utilisation[19].

We used the following datasets: METABRIC[70] (RNA-seq and mutations), TCGA-BRCA and TCGA-LUAD[71] (only RNA-seq data for both). We initially trained our sTabNet model with the above-defined attention mechanisms on METABRIC, testing different activation functions and comparing it with XGBoost, a fully connected neural network model (FCNN) and a CNN model. We then tested its in-domain and out-of-domain capabilities with and without fine-tuning. For the METABRIC dataset, we used XGBoost with a multiclass objective as suggested in the official library. The sTabNet was built as before, but using Gene Ontology[72] to define the sparse matrix $\boldsymbol{A}$ for the sparse layer. Given the more complex nature of the dataset, we added an extra layer at the end to increase the complexity of this sTabNet model.

We used dropout as regularisation. The sTabNet model was trained for 20 epochs and a batch size of 1024. We then used a 100-fold split for evaluation. Disease enrichment was conducted on the top 100 genes according to attention importance (ranked average of the 100 experiments). We used the GSEApy Python package and DisGeNet's disease database. To show the foundational capabilities of our sTabNet architecture, a sTabNet model was pre-trained on the METABRIC dataset to be used with and without fine-tuning. For feature extraction, we removed the last two layers from the frozen models and extracted the features for the TCGA-BRCA and TCGA-LUAD datasets. On the extracted features, we built a simple logistic regression model for a binary classification task. For the fine-tuning protocol, we added two fully connected layers to the frozen model. The model was fine-tuned using binary cross-entropy as a loss function. The plots show the results for 10-fold cross-validation.

Sparse net application to single-cell RNA-seq dataset

We used single-cell data from the Tumour Immune Single-cell Hub 2 (TISCH2) of breast cancer (GSE161529). We used annotated data with minimal preprocessing as suggested by ScanPy[73]. We used the sTabNet architecture as above to train two fresh networks for binary classification (tumour/normal cell) and multi-classification tasks (cellular type prediction). sTabNet was built for the METABRIC dataset, and we used XGBoost, as described above. The sTabNet model was trained for 20 epochs and a batch size of 1024. We then used a 10-fold random split for evaluation (a 10-fold split for evaluation was also used for XGBoost). We used a typical binary and multiclass final layer, i.e. for binary classification, we used a sigmoid final layer and binary cross-entropy as a loss, while for multiclass classification, we used a softmax layer and categorical cross-entropy.

Sparse net application to survival analysis

For survival analysis, we used the METABRIC dataset as described above, and we used the associated metadata for overall survival. We modified the sTabNet to adapt to the survival prediction task. Specifically, we used the

Breslow approximation as a loss function[74] for the last layer and measured the performance with a concordance index. We compare with the Scikit-survival library[46], which implements different algorithms for survival analysis. For this purpose, we selected two different gradient-boosted models (Gradient Boost and Component Wise Gradient Boost) and the fast extension of the support vector machine for survival analysis. For each algorithm, we conducted a 10-fold random split validation.

Generalisation of sTabNet for tabular datasets with no domain knowledge

In biological terms, a pathway is an approximation of locally connected features that interact among themselves to fulfil a biological function[16]. These could be seen as a group of connected nodes in the feature graph (e.g. a group of proteins in the protein-protein interaction graph). In this section, we will deal with the more general case when we do not have external knowledge of feature interactions, which is the case in most tabular datasets domains. We wanted to extend sTabNet to other domains where we need to enforce sparsity, but there is no knowledge available about the features or their interactions. We, therefore, hypothesised that this structure could be approximated as a random walk starting from each feature and walking in the feature graph, using Node2vec biased random walk[75]. In this way, we can explore the locality of each feature and its interactions.

Let $\boldsymbol{X} \in \mathbb{R}^{m\times n}$ be an input dataset. We define the feature graph G and its adjacency matrix $\boldsymbol{M} \in \mathbb{R}^{n\times n}$ in the following way. The nodes in G represent $\boldsymbol{X}$'s features, and the edges exist if there is a similarity between two features $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, n\}$. In other words, if the cosine similarity between the features i and j in $\boldsymbol{X}$ is higher than 0.5 or less than -0.5, $M_{ij} = 1$, otherwise $M_{ij} = 0$. The cosine similarity between two features is calculated by treating each feature as a vector of its values across all data instances. We then perform r Node2vec random walks of t steps for each node in the graph G (we did a parametric search for r and t and found that three and five are the best values, respectively). We considered a matrix $\boldsymbol{A} \in \{0, 1\}^{n\times r}$, where n is the dataset features and r is the number of random walks. For the random walk j if the node i is present in the random walk j, $A_{ij} = 1$, otherwise $A_{ij} = 0$. This process is described in Fig. 1C, D and Algorithm 1.

Algorithm 1. Generate graph and perform random walks.

Input : $X \in \mathbb{R}^{m\times n}$: a dataset with m samples and n features
n: number of features
r: number of random walks per node
t: length of each random walk
Output: $A \in \{0, 1\}^{n\times(r\cdot n)}$: matrix of random walks

Calculate cosine distance matrix $D \in \mathbb{R}^{n\times n}$ for all features;
Initialize an empty graph G;
for $i \leftarrow 1$ **to** n **do**
  for $j \leftarrow 1$ **to** n **do**
    if $i \neq j$ ***and*** $(D[i, j] > 0.5$ ***or*** $D[i, j] < -0.5)$ **then**
      Add an edge between nodes i and j in G;
    end
  end
end
Initialize matrix $A \in \{0, 1\}^{n\times(r\cdot n)}$ with zeros;
for $i \leftarrow 1$ **to** n **do**
  for $a \leftarrow 1$ **to** r **do**
    $v \leftarrow$ node i in G;
    $path \leftarrow$ Node2Vec RandomWalk$(v, \text{maxLength} = t)$;
    foreach $w \in path$ **do**
      Let $j \leftarrow$ index of node w in G;
      Set $A[j, (i - 1) \cdot r + a] \leftarrow 1$;
    end
  end
end

In a classical fully connected FFNN, each neuron can be considered a walk that connects this neuron to each node in the feature graph (Fig. 1C), thus learning a *global* approximation for the whole feature graph. Instead, a node in the sparse layer of our sTabNet, obtained by a random walk, is connected to relevant nodes and thus is learning a *local* approximation for the feature graph. Moreover, since we are using Node2vec, we can tune and control how big its locality is and how much of the locality of a feature we want to explore. Each neuron of the network specialises in a certain set of features (locality of the feature graph). This process reduces the impact of noise and irrelevant features. In fact, this enforced sparsity is making the model more robust to learning spurious correlations, which is a main cause of overfitting.

We also note that, while simple similarity-based clustering (e.g. cosine + k-means) can identify first-order relationships among features, it fails to capture multi-hop contextual dependencies that often underlie biological interactions. Node2Vec provides a principled way to encode these higher-order associations through random walks, yielding feature embeddings that preserve both local and global structure. This richer representation leads to more biologically coherent feature neighbourhoods and, consequently, more meaningful sparse connectivity patterns. To test our hypothesis that a random walk is a good approximation for domain knowledge connectivity, we used the benchmark datasets described by ref. 7 and focused on datasets with more than 20 features and fewer than 100 K examples. We built the sparse matrix A as described above; we performed three random walks of size 5 for each feature. We compared the sTabNet accuracy with XGBoost and logistic regression. Figure 4 shows the result of the training.

We also performed an ablation study. We added another attention layer *after* the sparse layer to study which neuron is more effective for our architecture. We extracted the attention weight (associated with each neuron) and conducted 100 different splits to account for the random walk variability, tracking which nodes were associated with the highest attention weight. We then removed these features, measured the performance and compared it with the model performance when we removed five features at random. The results are shown in Fig. 4E-G, where we can see that the performance deteriorated more when random walk features were removed, demonstrating that these random walk-identified features are more important than other features. In other words, our study shows that the random walk process is effective in identifying the most important connections between nodes and features, making our architecture well-suited to the problem at hand without using domain knowledge.

Data availability

No unique data were used in this study. All the data used are either published or generated by simulation. The process is based on a standard Python library (scikit-learn) for simulated data and is described in detail in the methods. For the other experiments, we used only published data, and the data are publicly accessible from the provided references.

Code availability

We are providing the code for the sparse neural network. As new implementations of the sTabNet become available, we will include them in the repository. Code can be accessed at: https://github.com/Tabular-Research-Group/sTabNet.

Received: 14 May 2025; Accepted: 22 November 2025;
Published online: 27 January 2026

References

1. Borisov, V. et al. Deep neural networks and tabular data: a survey. *IEEE Trans. Neural Netw. Learn. Syst.* **35**, 7499–7519 (2022).
2. Raieli, S. Tabular deep learning: a survey from small neural networks to large language models. Preprint at https://doi.org/10.36227/techrxiv.175753732.26052568/v1 (2025).
3. Saupin, G. *Practical Gradient Boosting—A deep dive into Gradient Boosting in Python* (AFNIL, 2023).
4. Noor, S. et al. Deep-m5U: a deep learning-based approach for RNA 5-methyluridine modification prediction using optimized feature integration. *BMC Bioinform.* **25**, 360 (2024).

5. Khan, S. et al. Xgboost-enhanced ensemble model using discriminative hybrid features for the prediction of sumoylation sites. *BioData Min.* **18**, 12 (2025).
6. Khan, S., AlQahtani, S. A., Noor, S. & Ahmad, N. Pssm-sumo: deep learning based intelligent model for prediction of sumoylation sites using discriminative features. *BMC Bioinform.* **25**, 284 (2024).
7. Grinsztajn, L., Oyallon, E. & Varoquaux, G. Why do tree-based models still outperform deep learning on typical tabular data? *Adv. Neural Inform. Proc. Sys.* **35**, 507–520 (2022).
8. Rubachev, I., Alekberov, A., Gorishniy, Y. & Babenko, A. Revisiting pretraining objectives for tabular deep learning. Preprint at https://doi.org/10.48550/arXiv.2207.03208 (2022).
9. Novakovsky, G., Dexter, N., Libbrecht, M. W., Wasserman, W. W. & Mostafavi, S. Obtaining genetics insights from deep learning via explainable artificial intelligence. *Nat. Rev. Genet.* **24**, 125–137 (2023).
10. Watson, D. S. et al. Clinical applications of machine learning algorithms: beyond the black box. *BMJ* **364**, l886 (2019).
11. Bayat, R., Pezeshki, M., Dohmatob, E., Lopez-Paz, D. & Vincent, P. The pitfalls of memorization: When memorization hurts generalization. Preprint at https://doi.org/10.48550/arXiv.2412.07684 (2024).
12. Roberts, M. et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nat. Mach. Intell.* **3**, 199–217 (2021).
13. Bommasani, R. et al. On the opportunities and risks of foundation models. Preprint at https://doi.org/10.48550/arXiv.2108.07258 (2022).
14. Frankle, J. & Carbin, M. The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks. In *Proc. Int. Conf. on Learning Representations* (ICLR) 2019. Available: https://openreview.net/pdf/2c35994ea2912e6517a87c50fc55faa58f0df150.pdf
15. Liu, S. & Wang, Z. Ten lessons we have learned in the new "sparseland": a short handbook for sparse neural network researchers. Preprint at https://doi.org/10.48550/arXiv.2302.02596 (2023).
16. Wysocka, M., Wysocki, O., Zufferey, M., Landers, D. & Freitas, A. A systematic review of biologically-informed deep learning models for cancer: fundamental trends for encoding and interpreting oncology data. *BMC Bioinform.* **24**, 198 (2023).
17. Surkov, A., Srinivas, V. & Gregorie, J. Unleashing the power of machine learning models in banking through explainable artificial intelligence (XAI). https://www2.deloitte.com/us/en/insights/industry/financial-services/explainable-ai-in-banking.html. Accessed 19 June 2023 (2022).
18. Elmarakeby, H. A. et al. Biologically informed deep neural network for prostate cancer classification and discovery. *Nature* **598**, 348–352 (2020).
19. Reel, P. S., Reel, S., Pearson, E., Trucco, E. & Jefferson, E. Using machine learning approaches for multi-omics data analysis: a review. *Biotechnol. Adv.* **49**, 107739 (2021).
20. López de Maturana, E. et al. Challenges in the integration of omics and non-omics data. *Genes* **10**, 238 (2019).
21. Vahabi, N. & Michailidis, G. Unsupervised multi-omics data integration methods: a comprehensive review. *Front. Genet.* **13**, 854752 (2022).
22. Attwaters, M. Bridging the multi-omics gap. *Nat. Rev. Genet.* **24**, 488 (2023).
23. Hao, J., Kim, Y., Kim, T.-K. & Kang, M. Pasnet: pathway-associated sparse deep neural network for prognosis prediction from high-throughput data. *BMC Bioinform.* **19**, 1–13 (2018).
24. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* **102**, 15545–15550 (2005).
25. Popov, S., Morozov, S. & Babenko, A. Neural Oblivious Decision Ensembles for Deep Learning on Tabular Data. In *Proc. Int. Conf. on Learning Representations* (ICLR) 2020. Available: https://openreview.net/pdf?id=r1eiu2VtwH.
26. Arik, S. Ö. & Pfister, T. TabNet: attentive interpretable tabular learning. In *Proc. AAAI Conference on Artificial Intelligence* Vol. 35, 6679–6687 https://doi.org/10.1609/aaai.v35i8.16826 (2021).
27. Shah, C., Du, Q. & Xu, Y. Enhanced tabnet: attentive interpretable tabular learning for hyperspectral image classification. *Remote Sens.* **14**, 716 (2022).
28. Whiteson, S., Stone, P., Stanley, K. O., Miikkulainen, R. & Kohl, N. Automatic feature selection in neuroevolution. In *Proc. 7th Annual Conference on Genetic and Evolutionary Computation* 1225–1232 https://doi.org/10.1145/1068009.1068210 (2005).
29. Grisci, B. I., Feltes, B. C. & Dorn, M. Neuroevolution as a tool for microarray gene expression pattern identification in cancer research. *J. Biomed. Inform.* **89**, 122–133 (2019).
30. Grisci, B. I., Krause, M. J. & Dorn, M. Relevance aggregation for neural networks interpretability and knowledge discovery on tabular data. *Inf. Sci.* **559**, 111–129 (2021).
31. Singh, A., Climente-González, H., Petrovich, N., Kawakami, E. & Yamada, M. FsNet: feature selection network on high-dimensional biological data. *Proc. Int. Joint Conf. Neural Netw*. 2023, 1–9 https://www.proceedings.com/content/069/069986webtoc.pdf (2023).
32. Gui, N., Ge, D. & Hu, Z. AFS: an attention-based mechanism for supervised feature selection. In *Proc. AAAI Conference on Artificial Intelligence* Vol. 33, 3705–3713 (2019).
33. Figueroa Barraza, J., López Droguett, E. & Martins, M. R. Towards interpretable deep learning: a feature selection framework for prognostics and health management using deep neural networks. *Sensors* **21**, 5888 (2021).
34. Lee, S. et al. Table2Image: interpretable tabular data classification with realistic image transformations. Preprint at https://doi.org/10.48550/arXiv.2412.06265 (2025).
35. Zhu, Y. et al. Converting tabular data into images for deep learning with convolutional neural networks. *Sci. Rep.* **11**, 11325 (2021).
36. Sharma, A., Vans, E., Shigemizu, D., Boroevich, K. A. & Tsunoda, T. DeepInsight: a methodology to transform a non-image data to an image for convolution neural network architecture. *Sci. Rep.* **9**, 11399 (2019).
37. Stuart, T. & Satija, R. Integrative single-cell analysis. *Nat. Rev. Genet.* **20**, 257–272 (2019).
38. Baysoy, A., Bai, Z., Satija, R. & Fan, R. The technological landscape and applications of single-cell multi-omics. *Nat. Rev. Mol. Cell Biol.* **24**, 695–713 (2023).
39. Lim, J. et al. Transitioning single-cell genomics into the clinic. *Nat. Rev. Genet.* **24**, 573–584 (2023).
40. Chattopadhyay, A. & Lu, T.-P. Gene-gene interaction: the curse of dimensionality. *Ann. Transl. Med.* **7**, 813 (2019).
41. George, B., Seals, S. & Aban, I. Survival analysis and regression models. *J. Nucl. Cardiol.* **21**, 686–694 (2014).
42. Wang, P., Li, Y. & Reddy, C. K. Machine learning for survival analysis: a survey. *ACM Comput. Surv. (CSUR)* **51**, 1–36 (2019).
43. Sarkar, K., Chowdhury, R. & Dasgupta, A. Analysis of survival data: challenges and algorithm-based model selection. *J. Clin. Diagn. Res.* **11**, LC14 (2017).
44. Cox, D. R. Regression models and life-tables. *J. R. Stat. Soc. Ser. B (Methodol.)* **34**, 187–202 (1972).
45. Mariani, L. et al. Prognostic factors for metachronous contralateral breast cancer: a comparison of the linear Cox regression model and its artificial neural network extension. *Breast Cancer Res. Treat.* **44**, 167–178 (1997).
46. Pölsterl, S. scikit-survival: A Library for time-to-event analysis built on top of scikit-learn. *J. Mach. Learn. Res.* **21**, 1–6 (2020).
47. Kadra, A., Lindauer, M., Hutter, F. & Grabocka, J. Well-tuned simple nets excel on tabular datasets. NeurIPS 2021, pp. 23928–23941

https://proceedings.neurips.cc/paper/2021/hash/c902b497eb972281fb5b4e206db38ee6-Abstract.html (2021).
48. Ulmer, D., Meijerink, L. & Cinà, G. Trust issues: uncertainty estimation does not enable reliable OOD detection on medical tabular data. ML4H-*NeurIPS Workshop Proc.* Vol. 136, 341–354 https://proceedings.mlr.press/v136/ulmer20a.html (2020).
49. Clements, J. M., Xu, D., Yousefi, N. & Efimov, D. Sequential deep learning for credit risk monitoring with tabular financial data. Preprint at https://doi.org/10.48550/arXiv.2012.15330 (2020).
50. Ahmed, M., Afzal, H., Majeed, A. & Khan, B. A survey of evolution in predictive models and impacting factors in customer churn. *Adv. Data Sci. Adapt. Anal.* **9**, 1750007 (2017).
51. Buczak, A. L. & Guven, E. A survey of data mining and machine learning methods for cyber security intrusion detection. *IEEE Commun. Surv. Tutor.* **18**, 1153–1176 (2015).
52. Urban, C. J. & Gates, K. M. Deep learning: a primer for psychologists. *Psychol. Methods* **26**, 743 (2021).
53. Jain, S. & Wallace, B. C. Attention is not Explanation. *Proc. EMNLP-IJCNLP 2019,* 11–20 https://www.sciencedirect.com/science/article/abs/pii/S1566253521002360?via%3Dihub (2019).
54. Jain, S. & Wallace, B. C. Attention is not explanation. In *Proc. 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3543–3556 (Association for Computational Linguistics, 2019).
55. Grimsley, C., Mayfield, E. & Bursten, J. R. S. Why attention is not explanation: surgical intervention and causal reasoning about neural models. In *Proc. Twelfth Language Resources and Evaluation Conference* (eds Calzolari, N. et al.) 1780–1790 (European Language Resources Association, 2020).
56. Bastings, J. & Filippova, K. The elephant in the interpretability room: Why use attention as explanation when we have saliency methods? *BlackboxNLP-EMNLP Workshop Proc.,* 149–155 https://aclanthology.org/2020.blackboxnlp-1.14.pdf (2020).
57. Grisci, B. I., Feltes, B. C., de Faria Poloni, J., Narloch, P. H. & Dorn, M. The use of gene expression datasets in feature selection research: 20 years of inherent bias? *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **14**, e1523 (2024).
58. Barbieri, M. C., Grisci, B. I. & Dorn, M. Analysis and comparison of feature selection methods towards performance and stability. *Expert Syst. Appl.* **249**, 123667 (2024).
59. Sidorenko, D. et al. Precious2gpt: the combination of multiomics pretrained transformer and conditional diffusion for artificial multi-omics multi-species multi-tissue sample generation. *npj Aging* **10**, 37 (2024).
60. Ooka, T. et al. Integrated-omics analysis with explainable deep networks on pathobiology of infant bronchiolitis. *npj Syst. Biol. Appl.* **10**, 93 (2024).
61. Yan, R., Islam, M. T. & Xing, L. Interpretable discovery of patterns in tabular data via spatially semantic topographic maps. *Nat. Biomed. Eng.* **9**, 471–482 (2025).
62. MA Basher, A. R., Hallinan, C. & Lee, K. Heterogeneity-preserving discriminative feature selection for disease-specific subtype discovery. *Nat. Commun.* **16**, 3593 (2025).
63. Bahdanau, D., Cho, K. & Bengio, Y. Neural machine translation by jointly learning to align and translate. *Proc. ICLR* https://www.semanticscholar.org/paper/Neural-Machine-Translation-by-Jointly-Learning-to-Bahdanau-Cho/fa72afa9b2cbc8f0d7b05d52548906610ffbb9c5 (2015).
64. Luong, T., Pham, H. & Manning, C. D. Effective approaches to attention-based neural machine translation. *Proc. EMNLP* 2015, 1412–1421 https://aclanthology.org/D15-1166/ (2015).
65. Mnih, V., Heess, N., Graves, A. & Kavukcuoglu, K. Recurrent models of visual attention. *Adv. Neural Inf. Process. Syst.* **27**, 2204–2212 https://proceedings.neurips.cc/paper_files/paper/2014/hash/3e456b31302cf8210edd4029292a40ad-Abstract.html (2014).
66. Vaswani, A. et al. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **30**, 5998–6008 https://papers.nips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html (2017).
67. Guyon, I. M. Design of experiments for the NIPS 2003 variable selection benchmark. Technical Report https://www.semanticscholar.org/paper/Design-of-experiments-for-the-NIPS-2003-variable-Guyon/b979fa88ca448fb08633f961131f45214b1cf109 (2003).
68. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. In *Proc. Advances in Neural Information Processing Systems* (eds Guyon, I. et al.) Vol. 30 (Curran Associates, Inc., 2017).
69. Tomsett, R., Harborne, D., Chakraborty, S., Gurram, P. & Preece, A. Sanity checks for saliency metrics. *AAAI Conf. Artif. Intell.* 34, 6021–6029 https://www.semanticscholar.org/paper/Sanity-Checks-for-Saliency-Metrics-Tomsett-Harborne/5be5f9660410c0aa23416ca005737861879c72dd (2020).
70. Curtis, C. et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346–352 (2012).
71. Network, T. C. G. A. R. et al. The cancer genome atlas pan-cancer analysis project. *Nat. Genet.* **45**, 1113–1120 (2013).
72. Ashburner, M. et al. Gene ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).
73. Wolf, F. A., Angerer, P. & Theis, F. J. Scanpy: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 1–5 (2018).
74. Yang X. et al. FastCPH: Efficient Survival Analysis for Neural Networks. Available here: https://arxiv.org/abs/2208.09793 (2022)
75. Grover, A. & Leskovec, J. node2vec: scalable feature learning for networks. *Proc. KDD 2016,* 855–864 https://dl.acm.org/doi/10.1145/2939672.2939754 (2016).

Acknowledgements

The authors want to thank Jonathan Schmiedt and Thomas Wursten, members of the OPM IT team, for their technical assistance during this project. The authors want to thank the anonymous reviewers for the insightful comments. S.R., N.J., S.V. and S.G. are employed by Oncodesign Precision Medicine and have not received additional funding.

Author contributions

SR and AA designed and conceptualised the study. AA supervised and verified the study and its concepts. SR performed the experiments, acquired data, and analysed the results. NJ provided datasets for the analysis. SR and AA wrote the manuscript. SV revised the manuscript. SG revised the manuscript and acquired funds. All authors contributed to the article and approved the submitted version.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s44387-025-00056-0.

Correspondence and requests for materials should be addressed to Salvatore Raieli or Abdulrahman Altahhan.