The Information Science Journal,
An Open Access Peer-Reviewed Journal

A Novel Machine Learning Approach for Data Analysis and Prediction

Elena V. Marchetti[1,2,#], Rajiv S. Bhatt[1,#], Thomas O. Nkemelu[1,2], Yuki Sato[1], Claire M. Dubois[1,2,*]

[1]Department of Data Science and Predictive Analytics, Northfield Institute of Technology, Hallberg, Sweden; [2]Centre for Applied Machine Learning Research, Northfield Institute of Technology, Hallberg, Sweden; Correspondence to: Dr. Claire M. Dubois, Department of Data Science and Predictive Analytics, Northfield Institute of Technology, Hallberg, Sweden; Email: c.dubois@northfieldit.example.edu

ABSTRACT

Accurate prediction from heterogeneous, high-dimensional data remains a central challenge in applied machine learning, particularly when datasets combine structured tabular records with time-dependent signals. This study proposes a hybrid convolutional-recurrent architecture, combining a convolutional feature extractor, a long short-term memory (LSTM) temporal branch, and an attention mechanism, for joint analysis and prediction across mixed data modalities. The proposed model was evaluated against three established baselines (logistic regression, random forest, and gradient boosting) on a combined structured and time-series benchmark of 48,600 records. The hybrid model achieved 94.7% accuracy and an F1-score of 0.938, outperforming the strongest baseline (gradient boosting) by 5.5 percentage points, at a modest increase in inference time (14.3 ms versus 9.8 ms). Ablation analysis showed that removing the attention layer reduced accuracy by 3.4 points and removing the LSTM branch reduced accuracy by 6.1 points, indicating that both temporal modeling and attention-based feature weighting contribute meaningfully to overall performance. These results suggest that hybrid architectures combining convolutional, recurrent, and attention components offer a practical and accurate approach for prediction tasks involving mixed structured and sequential data.

Keywords

Machine learning · Data analysis · Predictive modeling · Deep learning · Attention mechanism · Hybrid neural architecture

INTRODUCTION

The growth of high-volume, multi-modal data across scientific, industrial, and commercial domains has intensified demand for predictive models capable of handling both structured tabular records and time-dependent sequences within a single analytical pipeline. Conventional approaches typically treat these data types separately, applying tree-based ensemble methods to tabular data and recurrent or convolutional architectures to sequential data, which can limit their ability to capture cross-modal interactions relevant to prediction accuracy.

Ensemble methods such as random forest and gradient boosting remain widely used for structured data due to their strong performance, interpretability, and comparatively low computational cost [2;5]. However, these methods generally lack a native mechanism for modeling temporal dependencies, which limits their effectiveness on tasks where sequential patterns carry predictive signal. Deep learning architectures, particularly those combining convolutional and recurrent components, have shown promise in capturing both local feature patterns and longer-range temporal dependencies, but their application to mixed structured and sequential datasets has been comparatively underexplored [7;9].

This study proposes a hybrid machine learning architecture designed to jointly process structured and time-series data for prediction tasks. The model combines a convolutional feature

© 2026 The Information Science Journal

extraction layer, an LSTM branch for temporal modeling, and an attention mechanism intended to weight the relative contribution of extracted features. The purpose of this study was to evaluate the predictive accuracy of the proposed architecture relative to established baseline models, and to quantify the individual contribution of its architectural components through ablation analysis.

METHODS

Dataset

A combined benchmark dataset of 48,600 records was assembled from structured tabular features and associated time-series sequences drawn from a publicly available operational monitoring corpus. Each record included 42 structured features and a corresponding 30-step time-series sequence. The dataset was partitioned into training (70%), validation (15%), and test (15%) subsets using stratified sampling to preserve class balance across the prediction target.

Model architecture

The proposed hybrid model consists of three components. A one-dimensional convolutional block first extracts local feature patterns from the time-series input across three convolutional layers with progressively increasing filter depth. The resulting feature maps are passed to a two-layer LSTM branch to capture temporal dependencies across the sequence. An attention mechanism is applied over the LSTM outputs to assign learned weights to time steps most relevant to the prediction target, and the attended representation is concatenated with the structured tabular features before being passed to a fully connected output layer.

Baseline models

Three baseline models were trained for comparison: logistic regression, random forest (500 estimators), and gradient boosting (XGBoost, 500 estimators, maximum depth 6). All baselines were trained exclusively on the structured tabular features, consistent with standard practice for these model classes, and hyperparameters were tuned using grid search on the validation set.

Training and evaluation

The hybrid model was trained using the Adam optimizer with an initial learning rate of 1e-3, a batch size of 128, and early stopping based on validation loss over a maximum of 100 epochs. Model performance was evaluated on the held-out test set using accuracy, F1-score, and mean inference time per record. An ablation study was conducted by systematically removing the attention layer, the LSTM branch, and the data augmentation procedure applied during training, to assess the contribution of each component.

RESULTS

Predictive performance

The proposed hybrid model achieved the highest performance among all evaluated approaches, with 94.7% accuracy and an F1-score of 0.938 on the held-out test set (Table 1). This represented an improvement of 5.5 percentage points in accuracy over the strongest baseline, gradient boosting (89.2%), and a substantially larger margin over random forest (85.9%) and logistic regression (78.4%). The improvement in predictive performance was accompanied by an increase in mean inference time, from 9.8 ms for gradient boosting to 14.3 ms for the hybrid model.

© 2026 The Information Science Journal

Model	Dataset	Accuracy (%)	F1-Score	Inference Time (ms)
Logistic Regression (baseline)	Structured Tabular	78.4	0.741	2.1
Random Forest	Structured Tabular	85.9	0.831	6.4
Gradient Boosting (XGBoost)	Structured Tabular	89.2	0.874	9.8
Hybrid CNN-LSTM (proposed)	Time-Series + Tabular	94.7	0.938	14.3

Table 1: Predictive performance of the proposed hybrid model compared with baseline models on the held-out test set.

Ablation analysis

Ablation experiments indicated that all three architectural components contributed to overall model performance (Table 2). Removal of the LSTM branch, leaving only the convolutional pathway, produced the largest performance decrease, reducing accuracy by 6.1 percentage points to 88.6%. Removal of the attention layer reduced accuracy by 3.4 points to 91.3%, and removal of the data augmentation procedure reduced accuracy by 4.2 points to 90.5%. These results suggest that temporal modeling via the LSTM branch contributed most substantially to predictive performance, while attention-based weighting and data augmentation provided smaller but consistent additional gains.

Ablation Configuration	Accuracy (%)	Δ from Full Model
Full proposed model	94.7	—
Without attention layer	91.3	-3.4
Without LSTM branch (CNN only)	88.6	-6.1
Without data augmentation	90.5	-4.2

Table 2: Ablation study isolating the contribution of individual architectural components to model accuracy.

DISCUSSION

The results of this study indicate that a hybrid architecture combining convolutional feature extraction, recurrent temporal modeling, and attention-based weighting can outperform established ensemble baselines on prediction tasks involving mixed structured and time-series data. The magnitude of improvement over gradient boosting, the strongest baseline evaluated, suggests that the temporal information captured by the LSTM branch carries predictive signal not accessible to models trained solely on structured tabular features.

The ablation analysis further clarifies the relative contribution of each architectural component. The comparatively large performance decrease observed when removing the LSTM branch indicates that temporal dependencies are a primary driver of the model's predictive advantage, while the attention mechanism appears to provide a secondary refinement by weighting the most informative time steps. The contribution of data augmentation, though smaller, was consistent with its role in improving generalization on the held-out test set.

The increase in inference time associated with the hybrid model, while modest in absolute terms, may be a relevant consideration for deployment in latency-sensitive settings. Organizations weighing predictive accuracy against inference speed should consider whether the observed accuracy gains justify the additional computational cost for their specific application. This study has several limitations, including reliance on a single benchmark dataset and a fixed train-validation-test split; future work should evaluate

© 2026 The Information Science Journal

the proposed architecture across additional domains and assess its robustness under distributional shift and varying data quality conditions.

CONCLUSION

This study presented a hybrid convolutional-recurrent-attention architecture for prediction tasks involving combined structured and time-series data, and demonstrated that it outperforms conventional ensemble baselines by a substantial margin. Ablation analysis confirmed that both temporal modeling and attention-based feature weighting contribute meaningfully to the observed performance gains. These findings support the use of hybrid deep learning architectures for prediction tasks where structured and sequential data must be analyzed jointly, while highlighting the trade-off between predictive accuracy and inference latency that should inform deployment decisions.

DECLARATIONS

Funding

This study was supported by the Northfield Institute of Technology Data Science Research Fund [NIT-DSRF-2026-021] and the Hallberg Regional Innovation Grant [HRIG-1183].

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The benchmark dataset and processed model outputs supporting the findings of this study are available from the corresponding author upon reasonable request.

Author Contributions

Conceptualization, Dubois C.M. and Marchetti E.V.; methodology, Bhatt R.S.; software, Nkemelu T.O.; formal analysis, Sato Y.; investigation, Marchetti E.V.; writing—original draft, Marchetti E.V.; writing—review and editing, Dubois C.M. and Bhatt R.S.; supervision, Dubois C.M. All authors have read and agreed to the published version of the manuscript.

REFERENCES

1. Marchetti E.V., Bhatt R.S. Hybrid deep learning architectures for structured and sequential data. J Digit Syst Res. 2025;13(1):15-33.

2. Nkemelu T.O., Sato Y. Ensemble methods revisited: performance and limitations on tabular benchmarks. Ann Comput Sci. 2024;7(4):201-219.

3. Dubois C.M., Marchetti E.V. Attention mechanisms in time-series prediction: a comparative study. Mach Learn Appl. 2025;9(2):77-95.

4. Bhatt R.S., Nkemelu T.O. Convolutional-recurrent networks for multi-modal data fusion. Comput Ind Syst. 2024;18(3):142-160.

5. Sato Y., Dubois C.M. Gradient boosting in practice: a survey of production deployments. Data Sci Rev. 2023;5(1):33-50.

© 2026 The Information Science Journal