

AI TECHNIQUES FOR SURVEY DATA QUALITY: TRANSFORMERS AND GANS

Simona CAFIERI^{a*}, Gianmarco BORRATA^b

^a Italian National Institute of Statistics, Naples, Italy

^b Federico II University, Naples, Italy.

Abstract

This study develops a comparative framework for missing data imputation in social surveys, integrating classical statistical methods, machine learning algorithms, and advanced deep learning methods, including Transformers and Generative Adversarial Networks (GANs). The analysis specifically addresses the imputation of categorical and ordinal variables, ensuring full alignment between imputation strategies and classification-based evaluation metrics, thereby overcoming a key limitation of existing approaches. An empirical evaluation on real data from the Italian “Aspects of Daily Life” (AVQ) survey—covering approximately 50,000 individuals and 735 variables—demonstrates that deep learning models, particularly Transformer-based architectures, consistently outperform traditional and machine learning methods in terms of predictive accuracy, generalization, and preservation of distributional properties. The findings show that attention-based models effectively capture complex dependencies in high-dimensional survey data and can improve the quality and reliability of official statistics.

Keywords: Missing Data Imputation, Official Statistics, Transformers, GANs, Survey Data Quality

JEL Classification: C45, C53, C55, C81

* Corresponding author, Simona Cafieri – cafiери@istat.it

1. Introduction

Missing data have been extensively studied across multiple disciplines due to their pervasive presence in empirical datasets and their impact on statistical inference (Tsikriktsis, 2005; Graham et al., 2007; Baraldi & Enders, 2010; De Leeuw, 2011).

In official statistics and large-scale survey processing, item non-response and partial responses critically compromise data quality, introduce systematic bias, and reduce statistical efficiency, thereby affecting the reliability of key indicators used for evidence-based policymaking. For National Statistical Institutes (NSIs), accurate and timely imputation of missing values is therefore a central task in producing high-quality statistical outputs.

Within survey methodology, missing value imputation (MVI) has evolved from classical statistical procedures to advanced computational approaches, driven by increasing data complexity, dimensionality, and computational resources. Despite this progress, many traditional methods adopted in official statistics still inadequately handle the categorical and ordinal nature of key survey variables, leading to inconsistencies between imputation strategies and evaluation frameworks.

This issue is particularly relevant in socio-economic and health-related surveys, where variables are predominantly discrete and ordinal (e.g., health status, socio-economic classifications, behavioural indicators). Treating such variables as continuous or ignoring their structure distorts marginal distributions, reduces interpretability, and biases downstream statistical indicators.

Conventional techniques—mean substitution, regression-based imputation, and related model-based approaches—remain widely used for their interpretability and simplicity. However, they rely on restrictive parametric assumptions and often fail to capture complex nonlinear relationships among variables while maintaining the original data distribution.

This study focuses on imputing categorical and ordinal variables in large-scale social surveys using methods designed to preserve their discrete structure. The analysis includes neural models with explicit mechanisms for categorical variable reconstruction, ensuring consistency with the nature of survey variables. Unlike approaches primarily designed for continuous data, this work proposes a framework that maintains consistency between modelling and evaluation stages.

The main contribution is twofold. First, it provides a systematic comparison of classical statistical methods, machine learning algorithms, and advanced deep learning architectures within a unified framework tailored to categorical survey data. Second, it highlights the ability of attention-based models to capture complex dependencies and preserve aggregate statistical coherence in high-dimensional settings typical of statistical production—an essential feature for producing reliable official statistics.

The remainder of the paper is structured as follows. Section 2 reviews the literature, emphasising limitations of traditional methods and recent advances in data-driven imputation. Section 3 presents the proposed methodological framework and models, describes the dataset and experimental design. Section 4 discusses empirical results. Section 5 concludes with implications for official statistics and future research directions.

2. Literature review

The treatment of missing data has long been a central topic in statistical methodology, with numerous reviews summarizing methodological developments and practical applications (Little & Rubin, 2019; Lin & Tsai, 2020). Classical approaches rely on well-established statistical frameworks, including listwise and pairwise

deletion, hot-deck imputation, regression-based imputation, and likelihood-based methods such as the Expectation-Maximization (EM) algorithm (Dempster et al., 1977). Multiple imputation has further strengthened this field by providing a principled framework to account for uncertainty associated with missing data mechanisms (Rubin, 1987; Harel & Zhou, 2007). Despite their solid theoretical grounding and interpretability, these approaches often rely on restrictive assumptions regarding linearity, distributional forms, and missingness mechanisms such as MCAR and MAR. Moreover, most existing studies focus on continuous data settings, while the imputation of categorical variables in official statistics remains underexplored. Recent developments in official statistics have increasingly explored machine learning methods for data editing and imputation; however, their integration into production systems remains limited due to concerns regarding interpretability, methodological transparency, and compliance with statistical standards.

Machine learning methods, including decision trees, random forests, gradient boosting, and k-nearest neighbours, have demonstrated strong predictive performance by relaxing parametric assumptions and modelling nonlinear relationships (García-Laencina et al., 2010; Sun et al., 2023). These approaches have been widely applied in high-dimensional data contexts, where missing data imputation plays a crucial role in preserving data structure and analytical validity (Aittokallio, 2010; Liew et al., 2011). Random Forest algorithms have shown strong performance across different domains due to their robustness and ability to handle complex data structures (Belgiu & Drăguț, 2016). Tree-based methods also show strong performance in missing data imputation due to their flexibility and robustness (Nikfalazar et al., 2020).

Neural architectures such as Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, and Gated Recurrent Units (GRU) have shown effectiveness in capturing latent dependencies in complex datasets. GRU architectures, in particular, have been widely studied due to their simplified gating mechanisms and computational efficiency compared to LSTM variants (Dey & Salem, 2017). These architectures are included primarily for comparative purposes, given their widespread adoption in deep learning applications. In addition, deep learning approaches based on autoencoders have been successfully applied to missing data imputation, leveraging representation learning to reconstruct incomplete observations (Gondara & Wang, 2017). However, their adoption in official statistics remains limited due to challenges related to interpretability, computational cost, and alignment with survey design principles.

Within this evolving landscape, Transformer-based architectures and Generative Adversarial Networks (GANs) represent two of the most promising advancements. Transformers leverage self-attention mechanisms to model global dependencies in high-dimensional tabular data, while GAN-based approaches, such as Generative Adversarial Imputation Networks (GAIN), employ adversarial learning to generate plausible values for missing observations (Yoon et al., 2018; Casella et al., 2024; Du et al., 2023).

This study contributes to the literature by providing a comparative evaluation of classical statistical methods, machine learning techniques, and advanced deep learning architectures for missing data imputation in large-scale social surveys, with particular emphasis on their applicability to official statistical production.

3. Methodology

This study adopts a comparative methodological framework specifically designed to address categorical and ordinal variables, enabling a consistent comparison across models using classification-based evaluation metrics.

3.1 Benchmark Models

To establish a robust and comparable baseline, classical statistical imputation techniques and standard machine learning models are included in the analysis. These approaches are widely adopted in official statistics due to their interpretability and operational simplicity.

Classical imputation methods include hot-deck imputation and Multiple Imputation by Chained Equations (MICE), which provide a probabilistic framework for handling missing data and are consistent with survey sampling theory.

In addition, several machine learning models are considered, including linear and logistic regression, k-nearest neighbours, support vector machines, decision trees, random forests, and gradient boosting (XGBoost). These models are implemented using standard configurations commonly adopted in tabular classification tasks to ensure a fair and consistent comparison across approaches.

3.2 Deep Learning Architectures

Deep learning models are employed to assess their capacity to capture complex nonlinear relationships in large-scale multivariate survey data. Although originally developed for different data modalities, these architectures are adapted to tabular survey settings through feature embedding and representation learning techniques.

A Multilayer Perceptron (MLP) is used as a baseline neural architecture, providing a flexible framework for approximating nonlinear functions in high-dimensional feature spaces. Recurrent neural networks, including Long Short-Term Memory (LSTM) networks (Hochreiter & Schmidhuber, 1997) and Gated Recurrent Units (GRU), are also considered to explore their ability to model latent dependencies among variables, even though the data do not exhibit an inherently sequential structure.

In addition, one-dimensional Convolutional Neural Networks (Conv1D) are included to evaluate their effectiveness in capturing local patterns within structured tabular representations. Transformer architectures are implemented using self-attention mechanisms, which enable the modelling of global dependencies without relying on sequential assumptions. In this context, a masked reconstruction strategy is adopted for the imputation task, making Transformers particularly suitable for large-scale survey production environments affected by missing data.

Finally, Generative Adversarial Networks (GANs) are implemented through the Generative Adversarial Imputation Networks (GAIN) framework, where a generator network imputes missing values and a discriminator learns to distinguish observed from imputed entries.

To handle mixed-type survey variables, categorical features are encoded using a one-hot representation. Continuous variables are generated directly by the generator, while categorical variables are reconstructed through a softmax layer producing class probability distributions. During training, a Gumbel-Softmax relaxation is used to enable backpropagation through categorical variables. At inference time, continuous outputs associated with categorical variables are explicitly discretized through an argmax mapping over the predicted class probabilities, ensuring that all imputations belong exclusively to the admissible set of original survey categories (Gumbel, 1954; Jang et al., 2017).

3.3 Imputation Strategy and Evaluation Metrics

The imputation framework distinguishes between deterministic statistical and machine learning approaches and deep learning approaches involving stochastic optimization procedures during training, particularly in Transformer- and GAN-based architectures.

Model performance is evaluated using classification-based metrics, including Accuracy, Precision, Recall, and F1-score, which are appropriate given the categorical and ordinal nature of the target variables and allow a consistent assessment of imputation quality across methods.

This unified evaluation framework enables a direct comparison of imputation accuracy across all methods while explicitly accounting for the preservation of categorical structures in the data.

3.4. Data

The empirical analysis is based on the Italian “Aspects of Daily Life” (AVQ) survey conducted annually by the Italian National Institute of Statistics (ISTAT) (ISTAT, 2022; ISTAT, 2023). The AVQ is a nationally representative survey covering approximately 20,000 households and 50,000 individuals per year.

The study uses data from the 2021 and 2022 waves, comprising 735 variables. The survey adopts a mixed-mode data collection approach (CAWI/PAPI). As is typical in large-scale social surveys, item non-response generates missing values that may affect the accuracy and reliability of statistical analyses.

Missing values are assumed to follow a Missing At Random (MAR) mechanism conditional on observed covariates, in line with standard practices in survey data analysis (Rubin, 1976). To enable a controlled evaluation of imputation performance, additional missingness is synthetically introduced under a Missing Completely At Random (MCAR) mechanism on complete observations. Synthetic missingness was introduced by randomly masking a predefined proportion of observed values in complete cases under an MCAR mechanism. The masking procedure was repeated across different random seeds to reduce sampling variability and improve the robustness of the comparative evaluation. This experimental setup enables a controlled assessment of model performance while preserving the original structure and distributional characteristics of the survey data. Such a dual design supports the evaluation of imputation methods under both realistic and controlled conditions, while preserving the original data structure for robustness analysis.

The empirical evaluation focuses on two variables:

- Body Mass Index (BMI);
- Self-perceived health status (SALUT).

Both BMI and SALUT are treated as ordered categorical variables with a finite and ordered set of classes. Although BMI is originally computed as a continuous variable from self-reported weight and height, in this study it is analysed as an ordinal variable based on standard WHO classifications (underweight, normal weight, overweight, obese). This transformation ensures consistency with the categorical nature of the imputation task. SALUT represents increasing levels of health deterioration in self-perceived health status (very good, good, fair, bad, very bad). This representation aligns the variable definition with the imputation and evaluation framework adopted in the study.

In addition to predictive accuracy, the analysis evaluates the ability of each imputation method to preserve the original distributional characteristics of the data.

4. Results and Discussion

This section presents the empirical results of the comparative analysis of imputation methods. Model performance is evaluated using classification-based metrics appropriate for categorical and ordinal variables. All outputs are mapped to valid categorical classes, ensuring comparability across deterministic and stochastic approaches, including Transformer- and GAN-based models. Results are reported for the 2022 wave, while the corresponding results for 2021 are provided in the Appendix due to their substantial similarity.

4.1 Legend of Model Abbreviations

To enhance readability, Table no. 1 summarises the abbreviations used throughout the analysis.

Table no. 1: Model Abbreviations

Abbreviation	Model Name
LR	Logistic Regression
KNN	k-Nearest Neighbours
SVM	Support Vector Machine
DT	Decision Tree
RF	Random Forest
XGB	Extreme Gradient Boosting
MLP	Multilayer Perceptron
LSTM	Long Short-Term Memory
GRU	Gated Recurrent Unit
CONV1D	Convolutional Neural Network (1-Dimensional)
GAIN	Generative Adversarial Imputation Networks
TR	Transformer

Source: Authors' elaboration

4.2 Results for BMI

Table no. 2 reports the performance of Transformer (TR) and Generative Adversarial Imputation Networks (GAIN) models in imputing the Body Mass Index (BMI) variable for 2022. The results of the other models are presented in Table no. 6 of the appendix.

Table no. 2: Metrics for BMI (2022)

MODEL	Train Acc	Test Acc	Train Prec	Test Prec	Train Rec	Test Rec	Train F1	Test F1
TR	0.82	0.8	0.83	0.81	0.82	0.8	0.82	0.8
GAIN	0.8	0.71	0.81	0.72	0.8	0.71	0.8	0.71

Source: Authors' elaboration

The Transformer model achieves the best performance across all evaluation metrics, with a test F1-score of 0.80 and a limited gap between training and test results. GAIN shows lower test performance (F1-score of 0.71), while traditional machine learning models such as SVM and XGBoost achieve intermediate results. Ensemble methods, particularly Random Forest, display high training accuracy but reduced test performance, indicating mild overfitting. The Decision Tree model exhibits substantial overfitting, with perfect training scores and significantly lower test performance. Deep learning models such as MLP, GRU, and Conv1D show balanced but comparatively weaker performance.

4.3 Results for SALUT

Table no. 3 reports the performance of the models in imputing the self-perceived health status (SALUT) variable for 2022. The results of the other models are presented in Table no. 7 of the appendix.

Table no. 3: Metrics for SALUT (2022)

MODEL	Train Acc	Test Acc	Train Prec	Test Prec	Train Rec	Test Rec	Train F1	Test F1
TR	0.85	0.82	0.86	0.83	0.85	0.82	0.85	0.82
GAIN	0.8	0.73	0.81	0.74	0.8	0.73	0.8	0.73

Source: Authors' elaboration

The results confirm the same overall pattern observed for BMI. The Transformer model again achieves the highest performance, with a test F1-score of 0.82 and a minimal generalization gap. GAIN shows moderate performance (F1-score of 0.73), while traditional machine learning models remain competitive but do not reach the performance of the Transformer. Decision Trees models again exhibit poor generalization, with strong overfitting effects. Other deep learning architectures show stable but less competitive results. Despite this, GAIN remains consistent with the categorical constraints imposed by the unified evaluation framework.

4.4 Discussion

The results show that Transformer-based models outperform all competing methods across both variables and survey waves, with improvements in F1-score typically ranging between 6 and 9 percentage points, compared to traditional machine learning models and GAIN. The magnitude of the performance gap, particularly in terms of F1-scores, highlights the practical relevance of these differences.

The superior performance of the Transformer model indicates that self-attention mechanisms effectively capture complex dependency structures in high-dimensional data without relying on predefined assumptions about variable relationships.

In contrast, the GAIN framework, while theoretically appealing due to its generative formulation, exhibits weaker generalization performance in this setting. All GAIN-based imputations were mapped to admissible categorical classes prior to evaluation. Therefore, the observed performance gap does not stem from invalid category assignments, but rather from known challenges associated with adversarial training in discrete settings, including instability, mode collapse, and difficulties in accurately approximating complex categorical distributions. These findings suggest that, while generative adversarial approaches can be adapted to categorical imputation tasks, their practical effectiveness remains sensitive to training stability and model specification, particularly in complex survey environments.

Traditional machine learning models, such as Support Vector Machines and ensemble methods, provide competitive but intermediate performance. Their results indicate a reasonable capacity to model nonlinear relationships, although they tend to suffer from mild overfitting or limited flexibility compared to Transformer-based architectures. Decision Trees, in particular, show clear limitations due to overfitting, highlighting their reduced suitability for complex imputation tasks in large-scale survey systems.

From a methodological perspective, the findings highlight the importance of combining predictive accuracy with the preservation of distributional properties. In the context of official statistics, the ability of an imputation method to reproduce the empirical distribution of the data is as important as its performance at the individual level, as distortions in marginal distributions may lead to biased aggregate estimates.

Improved imputation quality may directly affect the estimation of key official statistics. More accurate reconstruction of BMI and self-perceived health categories can reduce distortions in prevalence estimates for obesity, overweight conditions, and poor health status. Since these indicators are widely used in public health monitoring and policy evaluation, preserving both individual-level accuracy and aggregate distributions is essential to ensure reliable evidence for decision-making processes. The empirical distribution analyses further support the suitability of Transformer-based imputations for survey applications.

Despite these promising results, some limitations should be acknowledged. The analysis is based on a specific survey (AVQ) and assumes a Missing At Random (MAR) mechanism conditional on observed covariates. Furthermore, the use of synthetically generated MCAR missingness, while useful for controlled evaluation, may not fully capture the complexity of real-world non-response mechanisms. Future research could extend the analysis to different surveys, alternative missingness mechanisms, and more complex data structures. In

addition, future work should further assess the computational cost and scalability of deep learning architectures in operational statistical production environments.

Overall, this study provides empirical evidence supporting the adoption of advanced AI techniques for imputation tasks in statistical outputs, particularly when dealing with multivariate survey categorical data. At the same time, the results underline the need for further research on adapting generative models to discrete data settings and improving their stability and generalization properties.

5. Conclusions

This study focuses on imputing discrete survey variables in official statistics using the ISTAT Aspects of Daily Life (AVQ) survey as a representative large-scale application. By comparing traditional statistical approaches, classical machine learning models, and modern deep learning architectures, the analysis provides a comprehensive assessment of imputation strategies for high-dimensional tabular survey data.

The results highlight the clear advantage of attention-based architectures for this task. Among the evaluated approaches, Transformer-based models proved the most robust and consistent across experimental settings, demonstrating superior performance in predictive accuracy while maintaining distributional consistency. These findings reinforce the need to move beyond traditional modelling assumptions when dealing with heterogeneous and complex survey environments.

A key contribution of this work lies in the adoption of an evaluation framework specifically tailored to categorical and ordinal variables, combining predictive and distributional perspectives. This approach aligns closely with the requirements of official statistics, where imputation quality must preserve both individual-level accuracy and statistical coherence.

These improvements are particularly relevant for large-scale survey production systems, where accurate imputation directly affects the reliability of aggregate health indicators.

From an operational perspective, the findings highlight the potential of advanced AI-based methods for modern statistical production processes. Improved imputation quality can enhance the reliability, coherence, and timeliness of official statistics, with direct implications for policy analysis and evidence-based decision making.

Despite these contributions, several limitations remain. The analysis is based on a specific survey context and a limited set of missing data scenarios. Future research should validate these findings across broader survey contexts and operational environments, and further investigate the role of explainability and hybrid statistical–AI approaches in operational environments.

Overall, the study highlights the potential of advanced AI methods to improve the quality, robustness, and operational reliability of modern official statistics

References

- Aittokallio, T. (2010). Dealing with missing values in large-scale studies: Microarray data imputation and beyond. *Briefings in Bioinformatics*, 11(2), 253–264. <https://doi.org/10.1093/bib/bbp060>
- Baraldi, A. N., & Enders, C. K. (2010). An introduction to modern missing data analyses. *Journal of School Psychology*, 48(1), 5–37. <https://doi.org/10.1016/j.jsp.2009.10.001>
- Belgiu, M., & Drăguț, L. (2016). Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing*, 114, 24–31. <https://doi.org/10.1016/j.isprsjprs.2016.01.011>

- Casella, M., Milano, N., Dolce, P., & Marocco, D. (2024). Transformers deep learning models for missing data imputation: An application of the ReMasker model on a psychometric scale. *Frontiers in Psychology*, 15, Article 1449272. <https://doi.org/10.3389/fpsyg.2024.1449272>
- De Leeuw, E. D. (2011). Reducing missing data in surveys: An overview of methods. *Quality & Quantity*, 45(1), 147–160. <https://doi.org/10.1007/s11135-010-9350-9>
- Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1), 1–38. <https://doi.org/10.1111/j.2517-6161.1977.tb01600.x>
- Dey, R., & Salem, F. M. (2017). Gate-variants of gated recurrent unit (GRU) neural networks. In *2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS)* (pp. 1597–1600). IEEE. <https://doi.org/10.1109/MWSCAS.2017.8053243>
- Du, T., Melis, L., & Wang, T. (2023). ReMasker: Imputing tabular data with masked autoencoding. *arXiv*. <https://arxiv.org/abs/2309.13793>
- García-Laencina, P. J., Sancho-Gómez, J. L., & Figueiras-Vidal, A. R. (2010). Pattern classification with missing data: A review. *Neural Computing and Applications*, 19, 263–282. <https://doi.org/10.1007/s00521-009-0295-6>
- Gondara, L., & Wang, K. (2017). Multiple imputation using deep denoising autoencoders. *arXiv*. <https://arxiv.org/abs/1705.02737>
- Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007). A gentle introduction to imputation of missing values. *Prevention Science*, 8(3), 206–213. <https://doi.org/10.1007/s11121-007-0099-9>
- Gumbel, E. J. (1954). *Statistical theory of extreme values and some practical applications*. National Bureau of Standards.
- Harel, O., & Zhou, X.-H. (2007). Multiple imputation: Review of theory, implementation, and software. *Statistics in Medicine*, 26(16), 3057–3077. <https://doi.org/10.1002/sim.2787>
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
- ISTAT. (2022). *Indagine Aspetti della vita quotidiana 2021* [Data set]. Istituto Nazionale di Statistica.
- ISTAT. (2023). *Indagine Aspetti della vita quotidiana 2022* [Data set]. Istituto Nazionale di Statistica.
- Jang, E., Gu, S., & Poole, B. (2017). Categorical reparameterization with Gumbel-Softmax. In *International Conference on Learning Representations*.
- Liew, A. W.-C., Law, N.-F., & Yan, H. (2011). Missing value imputation for gene expression data: Computational techniques to recover missing data from available information. *Briefings in Bioinformatics*, 12(5), 498–513. <https://doi.org/10.1093/bib/bbq080>
- Lin, W.-C., & Tsai, C.-F. (2020). Missing value imputation: A review and analysis of the literature. *Artificial Intelligence Review*, 53, 1487–1509. <https://doi.org/10.1007/s10462-019-09709-4>
- Little, R. J. A., & Rubin, D. B. (2019). *Statistical analysis with missing data* (3rd ed.). Wiley.
- Nikfalazar, S., Yeh, C.-H., Bedingfield, S., & Khorshidi, H. A. (2020). Missing data imputation using decision trees and fuzzy clustering with iterative learning. *Knowledge and Information Systems*, 62, 2419–2437. <https://doi.org/10.1007/s10115-020-01469-x>
- Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3), 581–592. <https://doi.org/10.1093/biomet/63.3.581>
- Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. Wiley.
- Sun, Y., Li, J., Xu, Y., Zhang, T., & Wang, X. (2023). Deep learning versus conventional methods for missing data imputation: A review and comparative study. *Expert Systems with Applications*, 227, Article 120201. <https://doi.org/10.1016/j.eswa.2023.120201>

- Tsikriktsis, N. (2005). A review of techniques for treating missing data in OM survey research. *Journal of Operations Management*, 24(1), 53–62. <https://doi.org/10.1016/j.jom.2005.03.001>
- Yoon, J., Jordon, J., & van der Schaar, M. (2018). GAIN: Missing data imputation using generative adversarial nets. In *Proceedings of the 35th International Conference on Machine Learning* (pp. 5689–5698). PMLR.

Appendix

Table no. 4: Metrics for BMI (2021)

MODEL	Train Acc	Test Acc	Train Prec	Test Prec	Train Rec	Test Rec	Train F1	Test F1
LR	0.7	0.69	0.71	0.69	0.7	0.68	0.7	0.68
GRU	0.75	0.7	0.76	0.71	0.75	0.7	0.75	0.7
CONV1D	0.73	0.69	0.74	0.7	0.73	0.69	0.73	0.69
MLP	0.74	0.69	0.75	0.7	0.74	0.69	0.74	0.69
SVM	0.78	0.71	0.79	0.72	0.78	0.71	0.78	0.71
KNN	0.74	0.68	0.75	0.69	0.74	0.68	0.74	0.68
XGBoost	0.77	0.7	0.78	0.71	0.77	0.7	0.77	0.7
RF	0.9	0.7	0.91	0.71	0.9	0.7	0.9	0.7
DT	1	0.5	1	0.52	1	0.5	1	0.5
TR	0.8	0.78	0.81	0.79	0.8	0.78	0.8	0.78
GAIN	0.81	0.72	0.82	0.73	0.81	0.72	0.81	0.72

Source: Authors' elaboration

Table no. 5: Metrics for SALUT (2021)

MODEL	Train Acc	Test Acc	Train Prec	Test Prec	Train Rec	Test Rec	Train F1	Test F1
LR	0.73	0.68	0.74	0.69	0.73	0.68	0.73	0.68
GRU	0.78	0.72	0.79	0.73	0.78	0.72	0.78	0.72
CONV1D	0.7	0.68	0.71	0.69	0.7	0.68	0.7	0.68
MLP	0.76	0.72	0.77	0.73	0.76	0.72	0.76	0.72
SVM	0.82	0.74	0.83	0.75	0.82	0.74	0.82	0.74
KNN	0.75	0.73	0.76	0.74	0.75	0.73	0.75	0.73
XGBoost	0.81	0.71	0.82	0.72	0.81	0.71	0.81	0.71
RF	0.91	0.7	0.92	0.71	0.91	0.7	0.91	0.7
DT	1	0.55	1	0.56	1	0.55	1	0.55
TR	0.83	0.81	0.84	0.82	0.83	0.81	0.83	0.81
GAIN	0.79	0.73	0.8	0.74	0.79	0.73	0.79	0.73

Source: Authors' elaboration

Table no. 6: Metrics for BMI (2022)

MODEL	Train Acc	Test Acc	Train Prec	Test Prec	Train Rec	Test Rec	Train F1	Test F1
LR	0.71	0.7	0.72	0.7	0.71	0.7	0.71	0.7
GRU	0.72	0.71	0.73	0.72	0.72	0.71	0.72	0.71
CONV1D	0.73	0.69	0.74	0.7	0.73	0.69	0.73	0.69
MLP	0.72	0.69	0.73	0.7	0.72	0.69	0.72	0.69
SVM	0.78	0.72	0.79	0.73	0.78	0.72	0.78	0.72
KNN	0.74	0.68	0.75	0.69	0.74	0.68	0.74	0.68
XGBoost	0.76	0.7	0.77	0.71	0.76	0.7	0.76	0.7
RF	0.92	0.69	0.93	0.7	0.92	0.69	0.92	0.69
DT	1	0.55	1	0.56	1	0.55	1	0.55

Source: Authors' elaboration

Table no. 7: Metrics for SALUT (2022)

MODEL	Train Acc	Test Acc	Train Prec	Test Prec	Train Rec	Test Rec	Train F1	Test F1
LR	0.72	0.67	0.73	0.68	0.72	0.67	0.72	0.67
GRU	0.75	0.71	0.76	0.72	0.75	0.71	0.75	0.71
CONV1D	0.73	0.7	0.74	0.71	0.73	0.7	0.73	0.7
MLP	0.75	0.72	0.76	0.73	0.75	0.72	0.75	0.72
SVM	0.78	0.73	0.79	0.74	0.78	0.73	0.78	0.73
KNN	0.77	0.74	0.78	0.75	0.77	0.74	0.77	0.74
XGBoost	0.8	0.72	0.81	0.73	0.8	0.72	0.8	0.72
RF	0.91	0.7	0.92	0.71	0.91	0.7	0.91	0.7
DT	1	0.52	1	0.53	1	0.52	1	0.52

Source: Authors' elaboration