Comparing Machine Learning and Deep Learning Approaches for ALS Diagnosis Using Blood-Derived Biomarkers and Synthetic Data

Ali Salman[1], Matteo Leoncini[1], Elena Niccolai[2], Jessica Mandrioli[3,4], Amedeo Amedei[2], and Ernesto Iadanza[1]
[1] *Department of Medical Biotechnologies - University of Siena, Siena, Italy,*
[2] *Department of Clinical and Experimental Medicine - University of Florence, Florence, Italy,*
[3] *Neurology Unit, Department of Neuroscience - Azienda Ospedaliero Universitaria Di Modena, Modena, Italy,*
[4] *Department of Biomedical, Metabolic and Neural Sciences - University of Modena and Reggio Emilia, Modena, Italy*

***Abstract*—Amyotrophic lateral sclerosis (ALS) is a relentlessly progressive neurodegenerative disease characterized by the loss of upper and lower motor neurons. Recent advances suggest that blood-derived biomarkers (BDBs) could enable earlier diagnosis and objective monitoring of disease progression. In this work, we assess the ability of both classical machine learning and deep neural networks to distinguish ALS patients from healthy controls (HC) using BDB measurements. To overcome the twin obstacles of small sample size and class imbalance in clinical datasets, we synthetically expanded our data by generating two balanced cohorts: one 10-fold and one 20-fold larger than half of the original dataset. Models were trained exclusively on these synthetic cohorts and then tested on the remaining 50% of the real data (60 records). Among all approaches, AdaBoost achieved the highest performance, with up to 96.7% accuracy, 0.9792 F1-score, and 0.9944 AUROC on the 20-fold set. The LGBMClassifier also performed strongly, particularly in precision and specificity (both 100%), while the deep learning model showed more variability between folds. Importantly, synthetic data preserves patient privacy while still enabling effective model development and evaluation. These findings demonstrate that carefully generated synthetic data can powerfully augment limited clinical datasets and pave the way for robust, blood-based computational diagnostics in ALS.**

***Keywords*—Amyotrophic Lateral Sclerosis, Blood-Derived Biomarkers, Synthetic Data Generation, Machine Learning.**

I. INTRODUCTION

Amyotrophic lateral sclerosis (ALS) is a neurodegenerative disease characterized by the progressive loss of both upper and lower motor neurons [1]. Following the onset of symptoms, the disease typically progresses with a survival time of 2 to 5 years [2]. Despite extensive research, the etiology of ALS remains largely unknown, and current therapeutic strategies are primarily focused on symptom management and extending survival [3]. Recent reviews highlight how artificial intelligence (AI) and microbiome research are converging to offer new paths for precision medicine and biomarker discovery [4]. ALS presents with considerable clinical heterogeneity, with spinal-onset being the most prevalent manifestation [5]. This variability has driven efforts to identify blood-derived biomarkers (BDBs) that can facilitate early diagnosis, improve disease monitoring, and provide insights beyond the neurodegenerative dimension of ALS.

Several BDBs have been evaluated for their diagnostic potential, including neurofilaments (NfL) [6], TAR DNA-binding protein 43 (TDP-43) [7], cerebrospinal fluid (CSF) chitinase [8], and cytokines [9]. A recent study by Niccolai et al. [10] explored the relationship between ALS and BDBs, focusing on microbiome molecules associated with metabolism and the immune system. This line of research builds upon previous work using AI to interpret gut microbiota patterns for clinical applications [11]. Their analysis also included Torque Teno Virus (TTV) viremia, which has been proposed as a potential biomarker for evaluating immune system functionality [12]. Their findings effectively discriminated ALS patients from healthy controls (HC) and identified distinct biological clusters among ALS patients.

Building upon this dataset, the objective of our study is to develop machine learning (ML) and deep learning (DL) models capable of accurately classifying ALS and HC individuals. A major challenge in this endeavor is the limited size and imbalance of the dataset, which could introduce biases in model training and evaluation. To address this, we utilized the Gaussian Copula Synthesizer (GCS) from the Synthetic Data Vault (SDV) Python package [13] to generate synthetic datasets. These synthetic datasets were used to train ML and DL algorithms. The trained models were subsequently tested on the real dataset, and their performance was assessed using a suite of evaluation metrics.

II. MATERIALS AND METHODS

This study addresses the classification of ALS and HC using ML and DL methods, supported by synthetic data generation. The methodology comprises four stages, as visually summarized in Figure 1.

A. Hardware and Software Resources

The study was conducted using an HP 470 G8 Notebook, with additional computational resources provided by Google

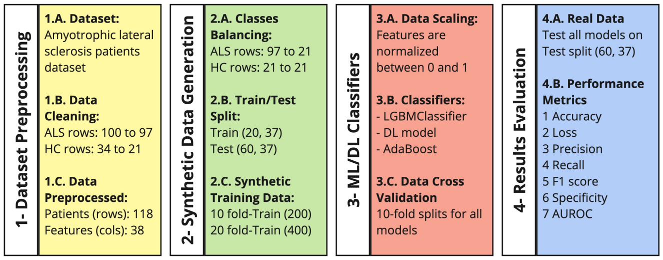

Fig. 1: Overview of the methodological framework: (1) Dataset preprocessing, (2) Synthetic data generation, (3) Machine learning and deep learning classifiers, and (4) Results evaluation.

Colab, a cloud-based Jupyter Notebook environment. All required Python libraries and packages were installed in this environment.

B. Dataset

The real dataset comprises 100 ALS patients and 34 HC participants, collected at the ALS Center of Modena University Hospital, Italy. BDBs were measured from serum samples, including cytokines (e.g., IL-1β, IL-6, TNF-α), short-chain fatty acids (e.g., acetic acid, propionic acid), and long-chain fatty acids (e.g., hexadecanoic acid). Additionally, Torque Teno Virus (TTV) viremia levels were analyzed.

C. Dataset Preprocessing

The preprocessing phase began with 38 features. One feature, patient ID, was excluded as it was irrelevant for synthetic data generation or the classification task. This resulted in a final feature set of 37 blood-derived biomarkers.

Next, dataset cleaning was performed to address missing values. Patient records containing "NULL" values were removed, reducing the dataset to 97 ALS patients and 21 HC. Synthetic data generation addressed the limited size and class imbalance in the cleaned dataset, creating larger and balanced datasets for subsequent model training and evaluation.

D. Synthetic Data Generation

Synthetic data was generated using the GCS to address the limited size and class imbalance in the dataset. Firstly, the real dataset was split in two dataframes: the train set (comprising of 48 ALS and 10 HC) and the test set (comprising of 49 ALS and 11 HC). Secondly, the train set has been undersampled to create a balanced subsample of 10 ALS and 10 HC, totalling 20 rows. Then it was used to generate two synthetic datasets: one with 200 rows (10 times the size of the subsample) and another with 400 rows (20 times the size of the subsample).

E. Classifiers

1) LightGBM Classifier: LightGBM is a gradient-boosting framework optimized for speed and efficiency. Its optimized decision tree algorithms and feature support make it well-suited for structured data classification, even with limited dataset sizes [14].

- *Classifier architecture*: Configured with 100 estimators, a learning rate of 0.1, and adaptive tree depth, it employs gradient-boosting techniques for classification.
- *Training*: LightGBM was trained on synthetic data using 10-fold stratified cross-validation. Features were standardized within each fold to prevent data leakage, and the binary cross-entropy loss function was used for optimization.

2) Deep Learning Classifier: A fully connected neural network was built using TensorFlow's Keras API [15].

- *Classifier architecture*: The model included dense layers with 96 and 32 units, ReLU activation, Batch Normalization, and Dropout layers to enhance generalization. The output layer used a sigmoid activation for binary classification.
- *Training*: The model employed stratified 10-fold cross-validation, with feature scaling applied separately to training and validation sets. Class weights were calculated to handle class imbalance. Early stopping and dynamic learning rate adjustment were used to prevent overfitting and improve convergence.

3) AdaBoost Classifier: AdaBoost, short for Adaptive Boosting, is an ensemble learning algorithm that combines weak classifiers to build a strong predictive model [16].

- *Classifier architecture*: Configured with 25 estimators and a learning rate of 0.1, it dynamically adjusted sample weights during training.
- *Training*: Synthetic data was used for 10-fold cross-validation, with features standardized within each fold. The model prioritized misclassified samples to enhance learning across iterations.

F. Code Availability

The code for this study is available at https://github.com/alexsalman/GNB2025.

III. RESULTS

A. Statistical Evaluation of Synthetic Data Generation

To assess the fidelity of the generated synthetic datasets, we used the "evaluate quality" function from the Synthetic Data Vault (SDV) library [17], which provides quantitative metrics for comparing synthetic data to the real dataset. Results are summarized in Table I and broken down as follows:

- *Column shapes*: This metric measures the similarity of marginal distributions for individual features between the real and synthetic datasets. Scores of approximately 78% for both the 10-fold and 20-fold synthetic datasets suggest a reasonable match in univariate distributions, meaning the synthetic data replicates the general spread and patterns of each individual feature.
- *Column pair trends*: This assesses how well bivariate relationships (e.g., correlations, trends) between feature pairs are preserved. Scores above 93% indicate that the synthetic data accurately captures inter-feature dependencies, which is critical for downstream tasks like

classification where model performance often hinges on these relationships.

- *Overall score*: This is a composite metric that aggregates the previous two, providing an overall measure of similarity between the real and synthetic datasets. With scores of 85.75% (10-fold) and 86.22% (20-fold), the synthetic datasets demonstrate strong fidelity, balancing realism with the privacy protection that comes from data synthesis.

These high similarity scores validate the structural quality of the synthetic datasets while inherently preserving patient privacy by design. As such, the generated data is not only safe for use in model development but also statistically representative of the original dataset, aligning with best practices for privacy-preserving data augmentation [18], [19].

TABLE I: Statistical similarities between real and synthetic ALS datasets, evaluated for two synthetic dataset sizes (10-fold and 20-fold the real dataset). Metrics include column shapes (marginal distributions), column pair trends (bivariate relationships), and overall score (aggregated similarity), highlighting the strong alignment between synthetic and real data.

Property	*Synthetic Data = 10-fold Real*	*Synthetic Data = 20-fold Real*
Column shapes	78.38%	78.86%
Column pair trends	93.13%	93.57%
Overall score	85.75%	86.22%

B. LightGBM Classifier

1) Training on 10-Fold-Real Synthetic Data: The LightGBM classifier demonstrated excellent performance, achieving an accuracy of 0.9000, F1 score of 0.9348, and AUROC of 0.9814. Both precision and specificity reached 1.0000, while recall was 0.8776, indicating strong generalization with a slight tendency to miss some ALS cases. The log loss was 0.2798, reflecting confident predictions.

2) Training on 20-Fold-Real Synthetic Data: Performance improved further with the larger synthetic dataset. Accuracy rose to 0.9167, and F1 score increased to 0.9462. Recall improved to 0.8980, while precision and specificity remained at 1.0000. The AUROC also increased to 0.9944, and log loss dropped to 0.1541, indicating highly confident and well-calibrated outputs.

C. Deep Learning Classifier

1) Training on 10-Fold-Real Synthetic Data: The deep learning model showed strong performance with 0.9000 accuracy, 0.9362 F1 score, and 0.9369 AUROC. Precision and recall were 0.9778 and 0.8980, respectively, while specificity was slightly lower at 0.9091. Log loss was 0.2694, suggesting reliable output probabilities.

2) Training on 20-Fold-Real Synthetic Data: Performance declined on the 20-fold set. Accuracy dropped to 0.7667, and the F1 score to 0.8333. Recall decreased sharply to 0.7143, although precision and specificity reached 1.0000, indicating the model became more conservative. AUROC was 0.9462, and log loss increased to 0.4954, reflecting overconfidence in some incorrect predictions.

D. AdaBoost Classifier

1) Training on 10-Fold-Real Synthetic Data: AdaBoost performed well with 0.9167 accuracy, 0.9462 F1 score, and 0.9889 AUROC. Precision, recall, and specificity were 1.0000, 0.8980, and 1.0000, respectively. Log loss was 0.3420, indicating solid model confidence and calibration.

2) Training on 20-Fold-Real Synthetic Data: Performance further improved across all metrics. Accuracy increased to 0.9667, F1 score to 0.9792, and AUROC to 0.9944. Precision and specificity remained perfect at 1.0000, and recall rose to 0.9592. Log loss dropped to 0.2852, showing the model made fewer and less confident mistakes.

E. Classifiers Comparison

In Table II is shown the performance metrics of the three classifiers, and below is a comparison of their performance:

1) LightGBM Classifier: LightGBM consistently delivered strong and stable results across both synthetic datasets. Its performance improved from the 10-fold to the 20-fold dataset, with accuracy rising from 0.9000 to 0.9167 and AUROC from 0.9814 to 0.9944. Notably, precision and specificity were perfect (1.0000) in both cases, and log loss decreased, reflecting confident and calibrated predictions. These trends demonstrate LightGBM's ability to scale effectively with more synthetic data, making it the most balanced and reliable model overall.

2) Deep Learning Classifier: The deep learning classifier showed strong performance on the 10-fold set (accuracy: 0.9000, F1: 0.9362, AUROC: 0.9369), but generalization degraded on the 20-fold dataset. Here, accuracy dropped to 0.7667, recall fell sharply to 0.7143, and log loss increased, indicating reduced confidence and potential overfitting to the synthetic distribution. Despite retaining perfect precision (1.0000) on the larger set, its lower recall points to missed ALS cases, limiting its diagnostic utility.

3) AdaBoost Classifier: AdaBoost demonstrated the most dramatic improvement with more data. From 0.9167 accuracy and 0.9889 AUROC on the 10-fold dataset, performance rose to 0.9667 accuracy and 0.9944 AUROC on the 20-fold set. Precision and specificity were consistently 1.0000, and recall increased from 0.8980 to 0.9592, while log loss dropped, suggesting both sensitivity and calibration improved with more training samples. AdaBoost rivaled LightGBM in several metrics but showed slightly higher log loss and more variability in earlier runs.

LightGBM classifier emerged as the most robust and consistent classifier, excelling across all metrics with excellent calibration, precision, and recall. AdaBoost also proved highly competitive, especially with larger synthetic datasets. The deep learning model, while initially strong, showed signs of overfitting and reduced recall, suggesting it may require further tuning or architectural adjustments to better handle synthetic clinical data.

TABLE II: Performance metrics of three classifiers (LightGBM, Deep Learning Model, and AdaBoost) evaluated on the test set. The classifiers were trained on synthetic datasets generated at 10x and 20x the size of the real dataset. Metrics include Accuracy, Log Loss, F1 Score, Precision, Recall, Specificity, and AUROC, highlighting the effectiveness of each model in classifying Amyotrophic Lateral Sclerosis (ALS) patients and healthy controls (HC).

Metrics	*LGBMClassifier*		*Deep Learning Model*		*AdaBoost*	
	10-fold Real	*20-fold Real*	*10-fold Real*	*20-fold Real*	*10-fold Real*	*20-fold Real*
Accuracy	0.9000	0.9167	0.9000	0.7667	0.9167	0.9667
Log Loss	0.2798	0.1541	0.2694	0.4954	0.3420	0.2852
F1 Score	0.9348	0.9462	0.9362	0.8333	0.9462	0.9792
Precision	1.0000	1.0000	0.9778	1.0000	1.0000	1.0000
Recall	0.8776	0.8980	0.8980	0.7143	0.8980	0.9592
Specificity	1.0000	1.0000	0.9091	1.0000	1.0000	1.0000
AUROC	0.9814	0.9944	0.9369	0.9462	0.9889	0.9944

IV. CONCLUSION

This study demonstrates the potential of synthetic data to train machine learning and deep learning models for distinguishing ALS patients from healthy controls using blood-derived biomarkers. By generating synthetic cohorts at 10x and 20x the size of the original dataset, we addressed challenges of small sample size, class imbalance, and data privacy, enabling robust model development without exposing sensitive patient information.

Among the evaluated classifiers, LightGBM consistently outperformed others, showing excellent accuracy, calibration, and generalization, especially with increased synthetic data. AdaBoost also performed strongly, particularly in recall and precision, while the deep learning model struggled to generalize on larger synthetic datasets.

These findings highlight synthetic data as a privacy-preserving and effective solution for enhancing machine learning in clinical settings. Future work should focus on improving synthetic data generation techniques and developing hybrid approaches that combine synthetic and real-world data to further strengthen model performance and generalizability.

REFERENCES

[1] Aline Furtado Bastos, Marco Orsini, Dionis Machado, Mariana Pimentel Mello, Sergio Nader, Júlio Guilherme Silva, Antonio M. da Silva Catharino, Marcos R.G. de Freitas, Alessandra Pereira, Luciane Lacerda Pessoa, Flavio R. Sztajnbok, Marco Araújo Leite, Osvaldo J.M. Nascimento, and Victor Hugo Bastos. Amyotrophic lateral sclerosis: One or multiple causes? *Neurology International*, 3(1), 2011.

[2] P. Masrori and P. Van Damme. Amyotrophic lateral sclerosis: a clinical review. *European Journal of Neurology*, 27(10):1918–1929, 2020.

[3] Lokesh Wijesekera and Nigel Leigh. Amyotrophic lateral sclerosis. *Orphanet Journal of Rare Diseases*, 4, 02 2009.

[4] Jasminka Hasic Telalovic, Serena Pillozzi, Rachele Fabbri, Alice Laffi, Daniele Lavacchi, Virginia Rossi, Lorenzo Dreoni, Francesca Spada, Nicola Fazio, Amedeo Amedei, et al. A machine learning decision support system (dss) for neuroendocrine tumor patients treated with somatostatin analog (ssa) therapy. *Diagnostics*, 11(5):804, 2021.

[5] P. Couratier, G. Lautrette, J.A. Luna, and P. Corcia. Phenotypic variability in amyotrophic lateral sclerosis. *Revue Neurologique*, 177(5):536–543, 2021. SFN 2020.

[6] Emily Feneberg, Patrick Oeckl, Petra Steinacker, Federico Verde, Christian Barro, Philip Van Damme, Elizabeth Gray, Julian Grosskreutz, Claude Jardel, Jens Kuhle, Sonja Koerner, Foudil Lamari, Maria del Mar Amador, Benjamin Mayer, Claudia Morelli, Petra Muckova, Susanne Petri, Koen Poesen, Joost Raaphorst, François Salachas, Vincenzo Silani, Beatrice Stubendorff, Martin R. Turner, Marcel M. Verbeek, Jochen H. Weishaupt, Patrick Weydt, Albert C. Ludolph, and Markus Otto. Multicenter evaluation of neurofilaments in early symptom onset amyotrophic lateral sclerosis. *Neurology*, 90(1):e22–e30, 2018.

[7] Marta Garcia Montojo, Sa Fa, Cyrus Rastegar, Elena Rita Simula, Tara Doucet-O'Hare, Yong-Han Cheng, Rachel Abrams, Nicholas Pasternack, Nasir Malik, Muzna Bachani, Brianna Disanza, Dragan Maric, Myoung-Hwa Lee, Herui Wang, Ulisses Santamaria, Wenxue Li, Kevon Sampson, Juan Lorenzo, Ignacio Sánchez, and Avindra Nath. Tdp-43 proteinopathy in als is triggered by loss of asrgl1 and associated with hml-2 expression. *Nature Communications*, 15, 05 2024.

[8] Alexander G Thompson, Elizabeth Gray, Alexander Bampton, Dominika Raciborska, Kevin Talbot, and Martin R Turner. Csf chitinase proteins in amyotrophic lateral sclerosis. *Journal of Neurology, Neurosurgery & Psychiatry*, 90(11):1215–1220, 2019.

[9] Zongzhi Jiang, Ziyi Wang, Xiaojing Wei, and Xue-Fan Yu. Inflammatory checkpoints in amyotrophic lateral sclerosis: From biomarkers to therapeutic targets. *Frontiers in Immunology*, 13, 2022.

[10] Elena Niccolai, Matteo Pedone, Ilaria Martinelli, Giulia Nannini, Simone Baldi, Cecilia Simonini, Leandro Di Gloria, Elisabetta Zucchi, Matteo Ramazzotti, Pietro Giorgio Spezia, Fabrizio Maggi, Gianluca Quaranta, Luca Masucci, Gianluca Bartolucci, Francesco Stingo, Jessica Mandrioli, and Amedeo Amedei. Amyotrophic lateral sclerosis stratification: unveiling patterns with virome, inflammation, and metabolism molecules. *Journal of Neurology*, 271, 04 2024.

[11] Ernesto Iadanza, Rachele Fabbri, Džana Bašić-ČiČak, Amedeo Amedei, and Jasminka Hasic Telalovic. Gut microbiota and artificial intelligence approaches: a scoping review. *Health and Technology*, 10(6):1343–1358, 2020.

[12] Omid Rezahosseini, Camilla Heldbjerg Drabe, Søren Schwartz Sørensen, Allan Rasmussen, Michael Perch, Sisse Rye Ostrowski, and Susanne Dam Nielsen. Torque-teno virus viral load as a potential endogenous marker of immune function in solid organ transplantation. *Transplantation Reviews*, 33(3):137–144, 2019.

[13] Neha Patki, Roy Wedge, and Kalyan Veeramachaneni. The synthetic data vault. In *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 399–410, 2016.

[14] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, volume 30, pages 3146–3154, 2017.

[15] François Chollet et al. Keras. https://keras.io, 2015. Accessed: 2025-01-17.

[16] Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. In *European Conference on Computational Learning Theory*, pages 23–37. Springer, 1995.

[17] Data Quality - Synthetic Data Vault. Data quality - synthetic data vault, 2023.

[18] Rémy Chapelle and Bruno Falissard. Statistical properties and privacy guarantees of an original distance-based fully synthetic data generation method. *arXiv preprint arXiv:2310.06571*, 2023.

[19] Hajra Murtaza, Musharif Ahmed, Naurin Farooq Khan, Ghulam Murtaza, Saad Zafar, and Ambreen Bano. Synthetic data generation: State of the art in health care domain. *Computer Science Review*, 48:100546, 2023.