Predicting human age from LCMS data using a fully connected neural network (FCNN) and feature selection with a sparse bilevel $\ell_{1,\infty}$ projection.

Nolwenn Peyratout [1,2],**Johan Lassen** [3], **Sonia Dagnino** [2], **Jorgen Hasserstrom** [3], **Palle Villensen** [3] **and Michel Barlaud** [1,*]

[1,*] *I3S Laboratory, CNRS, Côte d'Azur University, Sophia Antipolis, France*
[2,*] *Transporters in Imaging and Radiotherapy in Oncology (TIRO), School of Medicine, Direction de la Recherche Fondamentale, Institut Joliot, CEA, Côte d'Azur University. Nice. France*
[3,*] *Bioinformatics Research Center, Aarhus University, Aarhus, Denmark*

Correspondence*:

ABSTRACT

High-dimensional data, frequently encountered in metabolomics studies, often presents challenges for traditional statistical analyses due to the "curse of dimensionality" and the presence of technical noise and batch effects. These issues are particularly relevant in research on aging, where identifying reliable biomarkers from complex metabolic profiles is crucial. This study focuses on predicting chronological age from a large dataset of over 8,000 blood samples, originally collected for toxicological screenings from individuals suspected of driving under the influence of drugs. This unique dataset, while offering a large sample size, presents inherent challenges due to variations in sample handling, storage, and laboratory protocols.

To address these challenges, we employ a fully connected neural network (FCNN) enhanced with feature selection using structured sparse $\ell_{1,\infty}$ projection. This approach aims to extract the most informative features from the high-dimensional data while mitigating the impact of noise and batch effects. Our results demonstrate that the proposed FCNN model achieves a RMSE of 5.66 ± 0.07 years with only 4,983 features in predicting age, outperforming a standard FCNN with an RMSE of 5.78 ± 0.07 years. In particular, we find that a subset of $2,694$ features, selected through $\ell_{1,\infty}$ projection, provides comparable predictive accuracy as 5.71 ± 0.07 to utilizing the full set of features. This finding underscores the effectiveness of our feature selection method in identifying the most important metabolic signals for age prediction.

Keywords: Regression, Predicting human age, Fully connected Neural Network, Bilevel $\ell_{1,\infty}$ projection

INTRODUCTION

The study of human aging has attracted significant attention due to its implications for the extension of healthy lifespan. High Resolution Liquid Chromatography-Mass Spectrometry (HRLCMS) has emerged as a pivotal tool in aging research, enabling detailed analysis of metabolites that reflect the biochemical state of an organism. HRLCMS is particularly valuable for its high sensitivity and specificity in detecting a wide range of metabolites, which makes it indispensable for metabolomics studies aimed at understanding the aging process Liu et al. (2023).

Recent advances have seen the integration of HRLCMS with machine learning (ML) techniques to develop accurate age-prediction models Reveglia et al. (2021). The ability to predict chronological age from metabolic data not only provides insights into the biological understanding of aging, but also holds the potential to identify individuals at risk of age-related diseases. For example, analyzing CSF samples from healthy adults revealed significant age-related changes in metabolites such as cysteine, pantothenic acid, and 5-hydroxyindoleacetic acid Liu et al. (2023). These findings suggest that metabolic dysregulation is a hallmark of aging and can be quantitatively assessed using HRLCMS.

The integration of LC-MS and ML has led to significant advancements in the field of aging research. Studies have demonstrated that ML models can predict chronological age with high precision using metabolic profiles. For example, a study using data from the China Health and Retirement Longitudinal Study applied several ML algorithms, including Gradient Boosting Regressor and Random Forest, to develop a biological age measure that was significantly associated with physical disability and mortality Cao et al. (2021). Another study highlighted the use of ML to identify metabolic biomarkers for Alzheimer's disease, showcasing the potential of these techniques in early disease detection and monitoring Reveglia et al. (2021).

Lassen et al. previously modeled chronological age based on HRLCMS data from routine toxicological screenings of blood samples Lassen et al. (2023). These samples, while they present challenges in terms of experimental control and potential biases, provide a unique opportunity to investigate aging patterns within a large and diverse population.

In this paper, we try to model chronological age using a specialized type of fully connected neural network (FCNN) with feature projections. We use the same training/test scheme as in the original study Lassen et al. (2023) and show how sparse projection in combination with fully connected neural networks increases the prediction accuracy of human chronological age.

METHOD: REGRESSION USING A FULLY CONNECTED NEURAL NETWORK WITH FEATURE SELECTION USING THE BILEVEL $\ell_{1,\infty}$ PROJECTION

Deep neural networks have proven their efficiency for classification and feature selection in many domains, and have also been applied to omics data analyses Truchi et al. (2024); Chardin et al. (2022); Lassen et al. (2023).
Let X be the concatenated raw data matrix $(n \times m)$ (n is the number of patients and m the number of metabolites). Y is the vector $(n \times 1)$ of the age of each patient. Let $\hat{Y}$ be the encoded latent matrix (1×1). W is the matrix of the weights of the linear fully connected neural network (FCNN).

Criterion

The goal is to compute the network weights, W minimizing the regression loss. Moreover, to perform feature selection, as large datasets often present a relatively small number of informative features, we also want to sparsify the network, following the work proposed in Barlaud and Guyard (2020). Thus, instead of the classical computationally expensive Lagrangian regularization approach Hastie et al. (2004), we propose to minimize the following constrained approach introduced in Barlaud et al. (2017) in our Fully Connected neural Network (FCNN):

$$Loss(W) = \phi(\hat{Y}, Y) \text{ s.t. } BP_{\eta}^{1,\infty}(W). \tag{1}$$

Where $\hat{Y}$ is the estimate age by the neural network, ϕ is the mean square error loss, and $BP_{1,\infty}$ is the bilevel $\ell_{1,\infty}$ projection Barlaud et al. (2024).

We compute the weights using gradient with Adam method Kingma and Ba (2015). Note that low values of η imply high sparsity of the network. We use the double descent algorithm Barlaud and Guyard (2021).

Feature selection using the bilevel $\ell_{1,\infty}$ projection Barlaud et al. (2024)

The $\ell_{1,\infty}$ projection is of particular interest because it is able to set a whole set of columns to zero Bejar et al. (2021); Perez et al. (2023), instead of spreading zeros as done by the ℓ_1 norm. This makes it particularly interesting for reducing computational cost. However, the complexity of this algorithm remains an issue. The time complexity of this algorithm is $\mathcal{O}\big(nm * \log(nm)\big)$ for a matrix in $\mathbb{R}^{n \times m}$. Note that the complexity of the algorithm Perez et al. (2023) is, $\mathcal{O}\big(nm + J * \log(nm)\big)$ where J is a term that tends to 0 when the sparsity is high and $n \times m$. when the complexity is low.

The detailed propositions and algorithms for three bilevel projections $\ell_{1,\infty}$, $\ell_{1,1}$ and $\ell_{1,2}$ were provided by Barlaud et al. in Barlaud et al. (2024). The complexity of the bilevel algorithm is only $\mathcal{O}\big(nm\big)$. The code is available online[1] Note that the bilevel $\ell_{1,1}$ projection was used in single cell classification and feature selection Truchi et al. (2024). We propose here to use the bilevel $\ell_{1,\infty}$ projection Barlaud et al. (2024).

An evaluation metric using the Wasserstein distance

RMSE and MAE are classical metrics for regression evaluation. Here, we introduce the Wasserstein distance (or Kantorovich–Rubinstein metric) as another approach for the evaluation of regression results. The optimal transport problem or earthmover's distance was first formalized by Gaspard Monge in 1781. The Wasserstein distance is a natural way to compare the probability distributions of two variables and has been extensively used in the last decade in many machine learning applications Courty et al. (2016); Cuturi and Peyré (2018)

[1] https://github.com/MichelBarlaud/SAE-Supervised-Autoencoder-Omics

RESULTS

Preprocessing of data

Rather than using the PCA as done in the original study Lassen et al. (2023), we used the Local Outlier Factor (LOF) developed by Scikit-learn [2]. This method is more robust for identifying outliers, helping to isolate samples that deviate significantly from the majority. We fine-tuned the parameter to achieve the best results using the train split of the data before removing outliers from the full dataset.

After outlier removal, we log-transformed the data followed by a scaling (mean=0, standard deviation=1). After the preprocessing feature and sample preselection, our dataset was composed of 8,038 features and 8,099 samples.

Performance estimation

We train and estimate performance using the classical cross-validation of 90% of the data ("train set"), 8,184 samples, and we use the remaining 10% of the data, 815 samples, as external validation ("Final test") (See Figure 1) and [3].

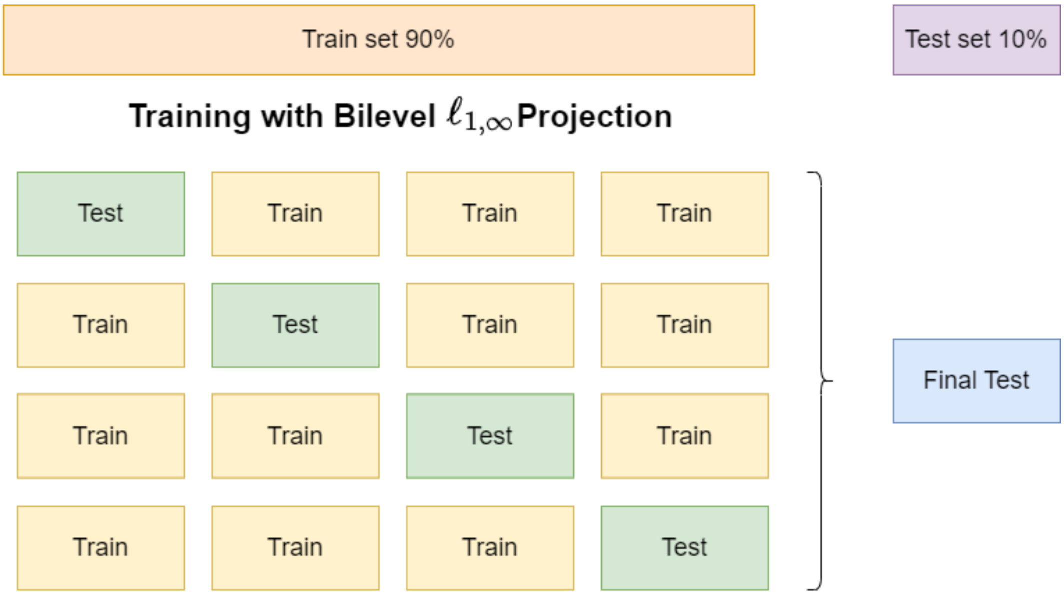

Figure 1. Train-Test-validation scheme

We train and estimate performance using the classical cross-validation of 90% of the data ("train set") and we use the remaining 10% of the data as external validation ("test set") (See Fig 1 [4]).

In the cross validation, we opted for a 4-fold cross validation, which means that we have 6,138 samples for the test and 2,046 samples for the train, each with 8,038 features. We trained a fully connected neural network using 2 seeds and the 4 folds. Testing on 2 seeds provided a more accurate overview of the model's statistical behavior, with all means and standard deviation computed over 8 folds.

[2] https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.LocalOutlierFactor.html

[3] https://github.com/NolwennPeyratout/FCNN-Age

[4] https://scikit-learn.org/stable/modules/cross_validation.html

During training, we carefully tuned the impact of each parameter on model performance, including the $SiLU$ activation continuous function, the batch size, and the learning rate. The best size of the three hidden layers of the fully connected neural network was set to $n = 300$ using cross validation.
Thus, the matrix modeling the connection between the first layer and the second layer has a size of $n = 300 \times m = 8038$. The feature selection is done with the $\ell_{1,\infty}$ projection applied to the first matrix. To remain consistent with this modification, we apply the projection on all the layers. We tune the parameter η of the projection in order to select features.
To avoid any leakage from test data to any test performance we split the data into a training and test split (9:1). All models were only using the training data to fit and evaluate model performance before finally being evaluated in the test set.
After initial outlier removal the dataset contained 8,099 samples with 8,038 features.

Model parameter influence on cross-validation performance

Using 4-fold cross-validation in the training data, we found the optimal number of features to be 5000 2 with an RMSE of 5.75 years. Evaluating performance in the test set resulted in the same general pattern, but an overall lower RMSE (5.66 at 5000 features).

Using mean absolute error gave slightly different results (3). While the cross-validation in the training data showed a minimal MAE at 5000 features, the test set showed a low MAE already at 2,500 features.

Metrics results of the CV test using our FCNN with the bilevel $\ell_{1,\infty}$ projection, as a function of the number of selected features. The line show the results using either cross-validation of the training set (blue) or test set (orange).

Figures 2 and 3 show that selecting only about 5,000 features ensure a RMSE of 5.75 years (cross validation) and an RMSE of 5.66 years (test set). For the MAE, we have a similar result, with 4.29 years using cross-validation and 4.25 with the test set. Surprisingly, the results of the test set showed a low MAE already at 2,500 features where the RMSE indicated 5,000 features to obtain the best prediction.

These loss distances curves, RMSE or MAE distance, as a function of the number of features, are convex. Therefore, this optimization requires a trade-off between error loss and the number of features. Note that it is the same trade-off to rate-distortion in lossy data compression Yochai and Michaeli (2019).

We also used an alternative metric, the Wasserstein distance between the true age distribution and the predicted age distribution. We compare it for several values of η, in order to find the best value. The theory is explained in . This metric measures the similarity between two distributions; in this case, we use it to assess the similarity between the true and predicted distributions. For our numerical evaluation, we use the metric provided by SciPy: [5].

The figure 4 shows that contrarily to previous RMSE and MAE curves, the Wasserstein distance provides an evident minimum for 2500 features for the cross-validation results and showed similar results for the test set.

[5] https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.wasserstein_distance.html

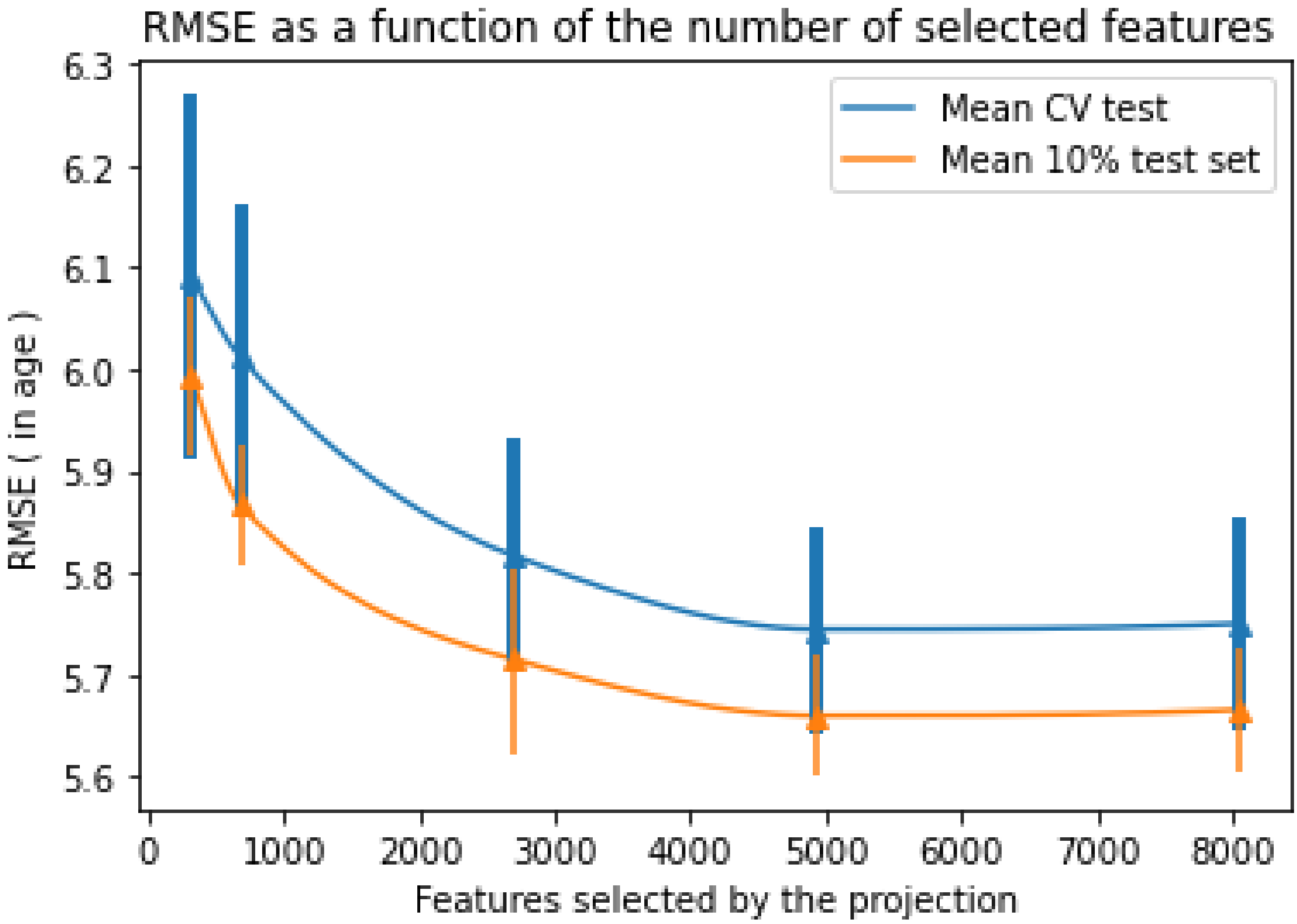

Figure 2. RMSE results

Thus, we conclude that using 2500 features is the best trade-off for RMSE and Wasserstein optimization. This conclusion is promising, indeed, we only need to compute the model with a third of the database to obtain good results. As a result, the computational cost of this learning is lower.

Figure 5 show the distribution of observed and predicted age from the cross validation results using 2,500 (A) and 5,000 (B) features. Both models underestimate the youngest and oldest samples. The dataset has an age bias, with a majority of samples being around 22 years old and a strong skew and results in a biased model that systematically underpredicts older samples.

The is also seen in figure 6 which reports the mean and standard deviation of predicted age during the final test, with 815 samples, with the best model of the cross validation. The color provides the sample size per year. First, we can not find any difference between 2,500 features and 5,000. Choosing only 2,500 features do not decrease the performance of the model. As discussed, the model has a bias of systematically predicting older individuals (over 30 years) to be younger than they are. This finding was also discovered with the study of the original paper, Lassen et al. (2023). Therefore, it could be explained by the fact that we have a clear sample's majority of 22 years old, and we also have few samples with more than 50 years old.

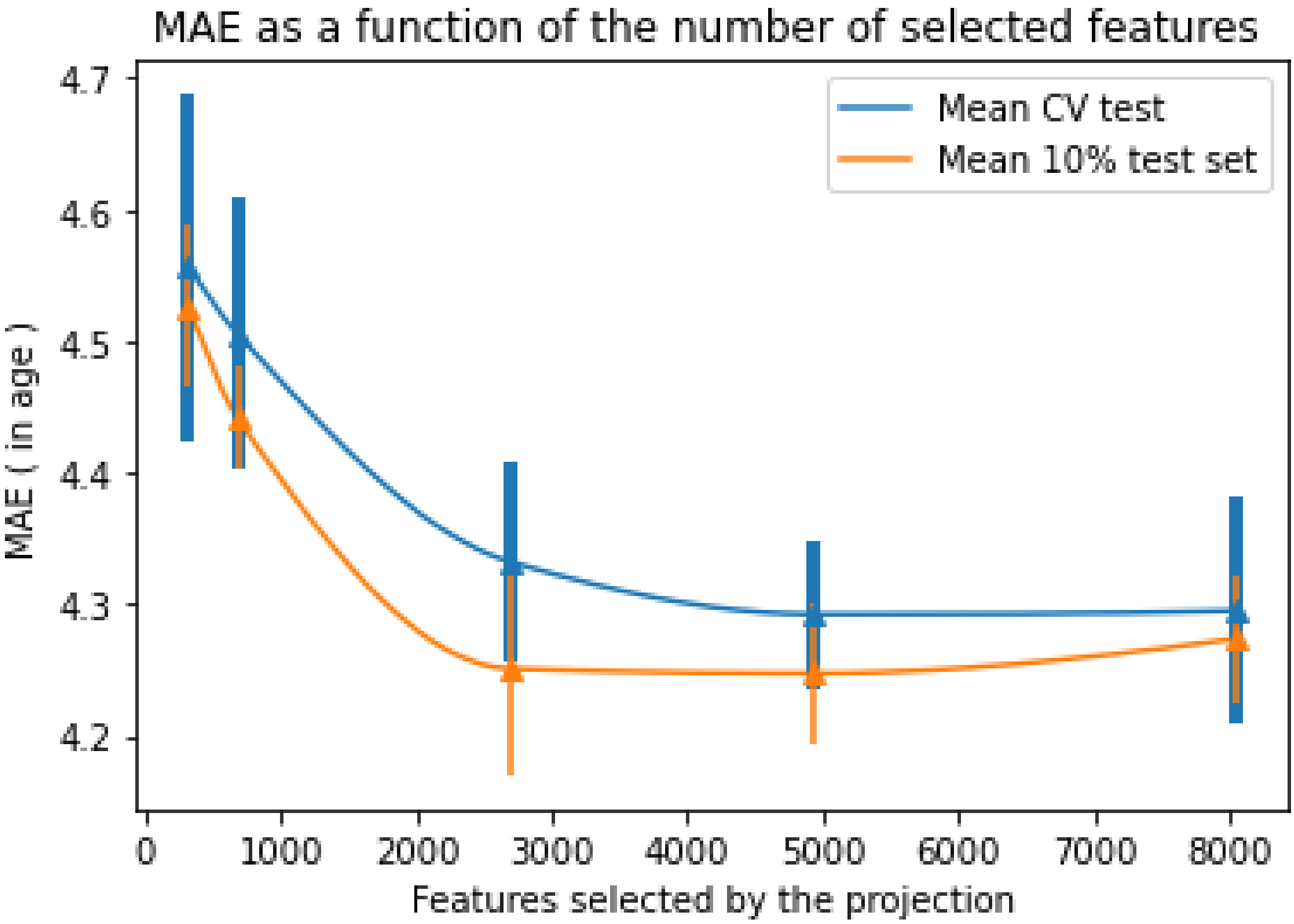

Figure 3. MAE results

A first approach to deal with this phenomenon could be to apply weights to very present samples to counterbalance the over-representation of young samples. In another point of view, this outcome is mathematically well known, and a possible solution to cope with this issue would be to minimize the Wasserstein distance Mohajerin Esfahani and Kuhn (2018). However, the minimization of the Wasserstein distance is still a hard topic out of the scope of this paper.

Projection comparison

Using two independent 4-fold cross-validations in the training set, we found that the bilevel FCNN method with projection outperformed the classical FCNN (without projection) across all metrics (Table 1) using both 2500 or 5000 features. Projection reduced the RMSE by 0.07 years when using 2,500 features compared to the classical method. Moreover, the bilevel projection with 2,500 features improved the Wasserstein Distance by 0.28 compared to the classical approach. This improvement applies not only to the performance, but also to the number of required features, as only 31% of the features are required. This reduction is significant for calculation costs, as it enables the gradient descent computation on 31% fewer neurons in the first layer.

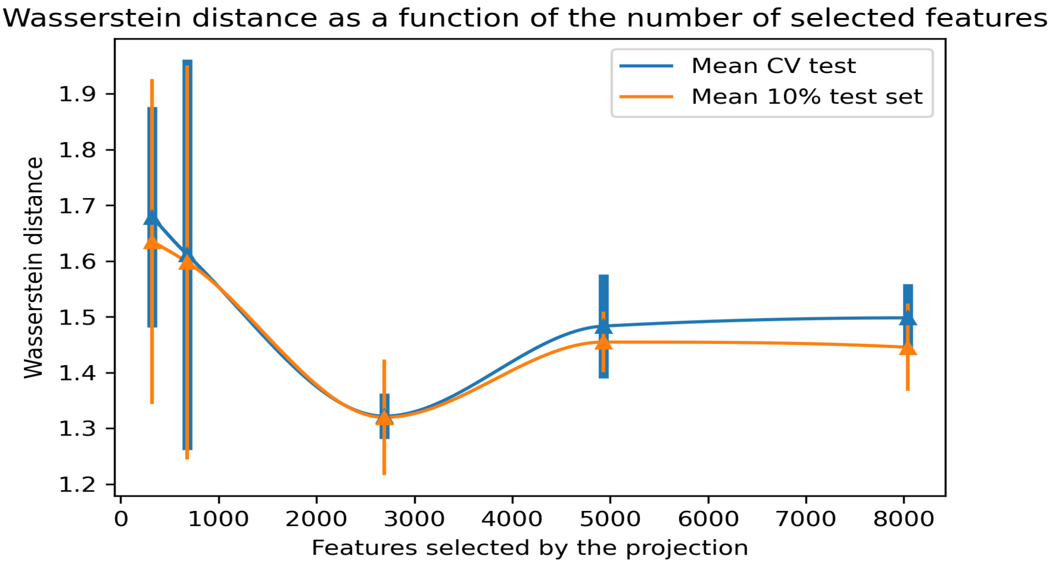

Figure 4. Wasserstein distance on the CV test and the 10% test using our FCNN with the bilevel $\ell_{1,\infty}$ projection, as a function of the number of selected features

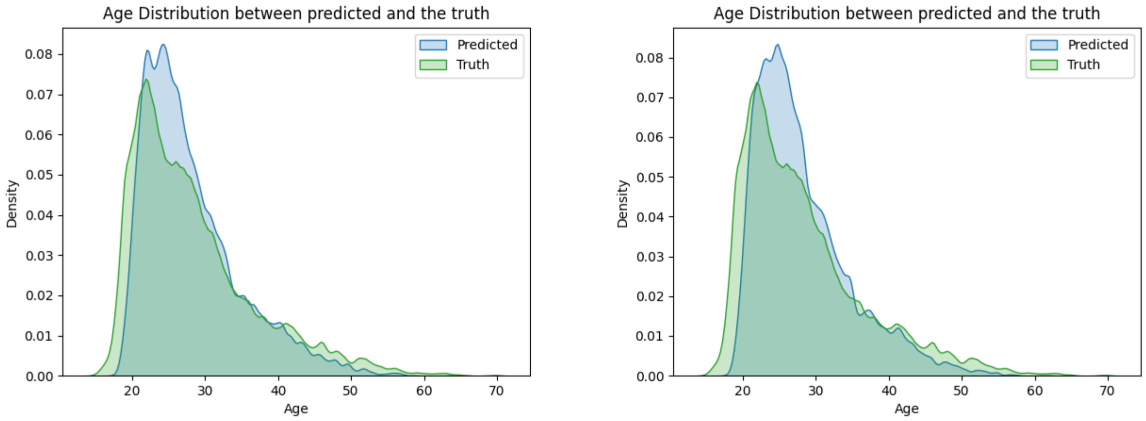

Figure 5. FCNN Bilevel distribution using a kernel method (bw=0.4) with 2500 and 5000 features of the Cross validation test set

	Mean RMSE CV test	Mean RMSE test	Mean WD CV test	Number of features
FCNN Bilevel $\ell_{1,\infty}$	5.81 ± 0.11	5.71 ± 0.09	1.32 ± 0.04	2,694
FCNN Bilevel $\ell_{1,\infty}$	5.75 ± 0.1	5.66 ± 0.06	1.48 ± 0.09	4,983
Classical FCNN	5.85 ± 0.09	5.78 ± 0.04	1.50 ± 0.06	8,038

Table 1. Train-validation test, RMSE and WD (Wasserstein distance); Comparison of methods and parameters for age estimation

In figure 7, we compare the predicted age as a function of the real age for both 2,500 selected features and the model without projection. We can not distinguish a difference between the two figures, suggesting that both neural networks exhibit similar bias for older ages. This observation implies that minimizing the

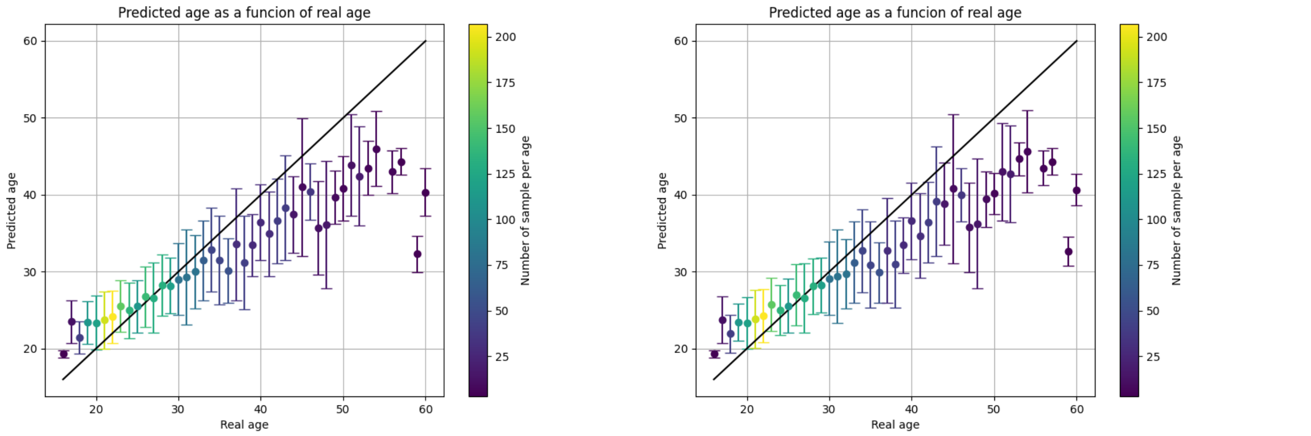

Figure 6. FCNN Bilevel distribution on the test set with the best fold. Left: using 2,500 features; Right: using 5,000 features

MSE may not solve the bias issue effectively. An alternative approach to solve this problem, could be to minimize the Wasserstein distance, since we have founded a clear optimal minimum.

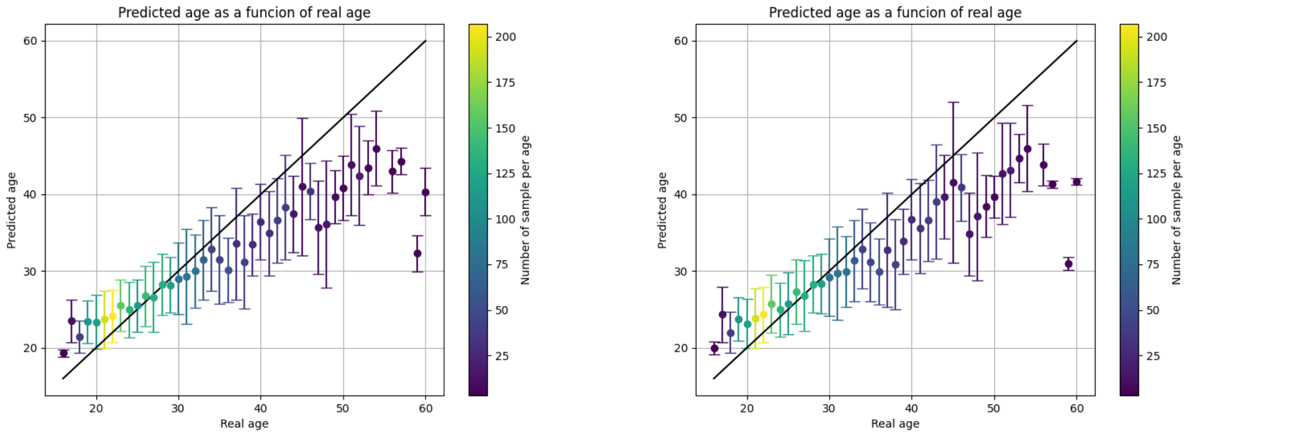

Figure 7. Distribution on the test set with the best fold. Left: with projection (2,500 features), Right without projection

Feature importance

The bilevel $\ell_{1\infty}$ projection is a structured projection, which means that certain feature weights are entirely set to zero. In figure 8 (left), the top fifteen features are ranked in descending order according to their normalized weights given by the Python library SHAP Lundberg and Lee (2017). This library computes the importance of each feature based on the learned weights of the neural network. We normalize these weights by the maximum value to determine the significance of each feature. We can distinguish a clear difference in feature's weight between the first and the tenth features for both figures, but we do not have a distinct break. Additionally, the curve flattens as features become less important, showing that the top features, though not a precise number, are predominant.

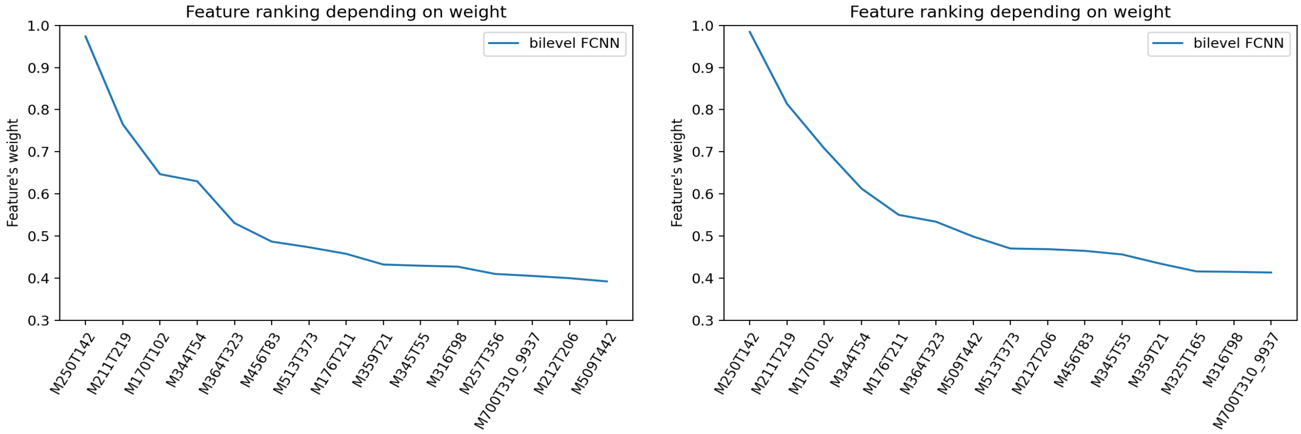

Figure 8. Features Ranking: Left for FCNN with 2500 features, Right FCNN with 5000 features

In figure 8 (right), features are normalized by the maximum value, as done previously. The ranked weights reveal the top discriminating metabolites, which can be interpreted as a perturbation signature. The major difference between the two figures is that, for the same top three features, the normalized weights given by SHAP for 2,500 features are slightly lower than those with 5,000 features, which may suggest as a less reliable top three. Note that the slope using bilevel $\ell_{1\infty}$ projection will give us a less flat curve compared to a classical deep neural network, resulting in a well-marked top features.

FCNN Bilevel 2500	FCNN Bilevel 5000	Original paper
M250T142	**M250T142**	**M250T142 [4-O-Dimethylallyl-tyrosine]**
M211T219	**M211T219**	**M211T219 [Cyclo(leu-pro)]**
M170T102	**M170T102**	**M170T102 [2,3-Dihydrodipicolinate]**
M344T54	M344T54	M255T346 [18-Nor-4(19),8,11,13-abietatetraene]
M364T323	**M176T211**	M260T236
M456T83	M364T323	M257T356
M513T373	M509T442	**M176T211**
M176T211	M513T373	M469T561
M359T21	M212T206	M521T504
M345T55	M456T83	M220T196

Table 2. Top 10 features in descending order of weight. Features found in across all three lists are highlighted in red. Features found in across all the first two are highlighted in blue.

To establish a more accurate comparison of the identified features, we constructed a table (Table 2) showing the top ten features discovered in our FCNN using 2,500 and 5,000 features, alongside those identified in the original study Lassen et al. (2023). The top three metabolites appear in identical ranks across both studies, meaning they converge on the same result and one additional feature (M176T211) is also shared across all three (highlighted in red). Three additional features (highlighted in blue) are shared between the two projection networks, showing the reliability of this approach with different value of η. Feature importance is very high for a few features, but decrease and flattens out really fast (figure 8). Many features will thus have similar importance (around 0.4) and may change rank between runs. It was only possible to annotate the first four features in the original paper.

DISCUSSION

In summary, we find that the $\ell_{1,\infty}$ projection improves prediction results and use fewer features than the original paper Lassen et al. (2023). The use of the $\ell_{1,\infty}$ reduces the number of features during learning and, consequently, the computational cost with no loss of performance for this dataset. The $\ell_{1,\infty}$ projection is particularly advantageous over the classical ℓ_1 projection, as it selects entire columns, and thus relevant features, rather than isolated points within the matrix. As a result, learning with the $\ell_{1,\infty}$ projection removes noisy features while improving RMSE, MAE and Wasserstein distance compared to the classical fully connected neural network.

Throughout this study, we observed a bias with younger samples being over-predicted and older samples being under-predicted. A possible extension of this study it to implement the Wasserstein distance as an alternative loss function for the network fitting. Here, we only employed the MSE loss and checked the Wasserstein distance as a performance metric when tuning the parameter η that controls the regularization. However, this topic involves complex optimal transport theory, which falls outside the scope of this paper.

The bilevel $\ell_{1,\infty}$ projection has already proved its efficiency in single cell application Truchi et al. (2024); Barlaud et al. (2024). In these case, the projection selected a limited number of selected features (hundreds) and provides a large accuracy improvement by 10% compared to standard network. Even though metabolomics and single cell gene expression data are quite different, our results show that the projection seem to be beneficial in both cases. This calls for further testing of the $\ell_{1,\infty}$ projection in other high-dimensional biomedical datasets, to see if in the projection approach generally performs better or on par with existing state-of-the-art methods.

According to the outcomes obtained with the RMSE and the Wasserstein distance in our metabolomic application, the $\ell_{1,\infty}$ projection provides a limited selected feature, around 30%, which correspond to 2,500 selected features. This led to moderate improvements in RMSE or MAE, but a more substantial improvement in the Wasserstein distance.

The features selection results should be interpreted with caution, in fact, the data is from drivers suspected of driving under the influence of drugs. The features found may therefore have been influenced by drugs intake and may only be relevant within the context of this dataset.

Samples

The dataset as described in Lassen et al. (2023) consist of blood samples collected from drivers suspected of drug-impaired driving between January 2017 and December 2020. The cohort is 93% male, with a mean age of 28.9 ± 9.2 years, and a skewed age distribution.

The dataset presents different challenges; the samples were not collected under controlled conditions ideal for metabolomics analysis. Variations in sample handling, storage times, and even changes in laboratory protocols, such as the switch from FC to FX sample tubes, introduce experimental noise and batch effects that can obscure true biological signals.

Data were fully anonymized prior to analysis. Untargeted metabolomics was performed with UHPLC-QTOF across 394 batches. Peak picking was performed with XCMS and allowed the identification of 12,686 features, excluding those with >20 %missing values per batch.

For further details on the LCMS details, please see Telving and Andreasen (2016).

Data declaration and availability

All methods were carried out in accordance with relevant guidelines and regulations. All experimental protocols were approved by relevant Danish authorities.

The data were provided by the Department of Forensic Medicine, Aarhus University but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of Department of Forensic Medicine, Aarhus University.

ACKNOWLEDGEMENTS

PV and JL are thankful to the Aarhus University research foundation and Innovation Fund Denmark for funding. NP is thankful to the Polytech department of the Côte d'Azur University for funding the internship and thankful to I3S laboratory for managing the grant. SD is thankful to the ANR Pre.S.Age for funding. All authors are thankful to the Department of Forensic Medicine, Aarhus University for allowing access to the data.

AUTHOR CONTRIBUTIONS

MB wrote the model section, NP and MB designed the pytorch code and the experiment. NP and JL performed data handling, analysis and conclusions. MB, PV, and SD supervised the project. All authors participated in approval of the manuscript.

ADDITIONAL INFORMATION

All authors declare no competing interests.

REFERENCES

Barlaud, M., Belhajali, W., Combettes, P., and Fillatre, L. (2017). Classification and regression using an outer approximation projection-gradient method. vol. 65, 4635–4643

Barlaud, M. and Guyard, F. (2020). Learning sparse deep neural networks using efficient structured projections on convex constraints for green ai. *International Conference on Pattern Recognition, Milan* , 1566–1573

Barlaud, M. and Guyard, F. (2021). Learning a sparse generative non-parametric supervised autoencoder. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing, Toronto, Canada*

Barlaud, M., Perez, G., and Marmorat, J.-P. (2024). Linear time bi-level l1,infini projection ; application to feature selection and sparsification of auto-encoders neural networks. *arXiv 2407.16293v1 [cs.LG]*

Bejar, B., Dokmanić, I., and Vidal, R. (2021). The fastest $\ell_{1,\infty}$ prox in the West. *IEEE transactions on pattern analysis and machine intelligence* 44, 3858–3869

Cao, X., Yang, G., Jin, X., He, L., Li, X., Zheng, Z., et al. (2021). A machine learning-based aging measure among middle-aged and older chinese adults: The china health and retirement longitudinal study. *Frontiers in Medicine, 8, 698851*

Chardin, D., Gille, C., Pourcher, T., Humbert, O., and Barlaud, M. (2022). Learning a confidence score and the latent space of a new supervised autoencoder for diagnosis and prognosis in clinical metabolomic studies. *BMC Bioinformatics* 23

Courty, N., Flamary, R., Tuia, D., and Rakotomamonjy, A. (2016). Optimal transport for domain adaptation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*

Cuturi, M. and Peyré, G. (2018). Semidual regularized optimal transport. *SIAM Review* 60, 941–965. doi:10.1137/18m1208654

Hastie, T., Rosset, S., Tibshirani, R., and Zhu, J. (2004). The entire regularization path for the support vector machine. *Journal of Machine Learning Research* 5, 1391–1415

Kingma, D. and Ba, J. (2015). a method for stochastic optimization. *International Conference on Learning Representations* , 1–13

Lassen, J. K., Wang, T., Nielsen, K. L., Hasselstrøm, J. B., and Johannsen, P., M.and Villesen (2023). Large-scale metabolomics: Predicting biological age using 10,133 routine untargeted lc–ms measurements. *Wiley, Aging Cell*

Liu, F.-C., Cheng, M.-L., Lo, C.-J., Hsu, W.-C., Lin, G., and Lin, H.-T. (2023). Exploring the aging process of cognitively healthy adults by analyzing cerebrospinal fluid metabolomics using liquid chromatography-tandem mass spectrometry. *BMC Geriatrics, 23, 217*

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Neural Information Processing Systems, Barcelone, Spain* 30

Mohajerin Esfahani, P. and Kuhn, D. (2018). Data-driven distributionally robust optimization using the wasserstein metric: performance guarantees and tractable reformulations. *Mathematical Programming* 171, 115–166

Perez, G., Condat, L., and Barlaud, M. (2023). Near-linear time projection onto the l1,infty ball application to sparse autoencoders. *IEEE International Conference on Tools with Artificial Intelligence Washigton USA 2024*

Reveglia, P., Paolillo, C., Ferretti, G., De Carlo, A., Angiolillo, A., Nasso, R., et al. (2021). Challenges in lc–ms-based metabolomics for alzheimer's disease early detection: targeted approaches versus untargeted approaches. *Metabolomics, 17, 78*

Telving, J. B., R.and Hasselstrøm and Andreasen, M. F. (2016). Targeted toxicological screening for acidic, neutral and basic substances in postmortem and antemortem whole blood using simple protein precipitation and uplc-hr-tof-ms. *Forensic Science International*

Truchi, M., Lacoux, C., Gille, C., Fassy, J., Magnone, V., Lopes Goncalves, R., et al. (2024). Detecting subtle transcriptomic perturbations induced by lncrnas knock-down in single-cell crispri screening using a new sparse supervised autoencoder neural network. *Frontiers in Bioinformatics*

Yochai, B. and Michaeli, T. (2019). Rethinking lossy compression: The rate-distortion-perception tradeoff. *International Conference on Machine Learning*