Evaluation of Deep Learning Strategies for Automated Cortisol Quantification in Hair Matrix via LC-MS/MS

Heitor de B. Santos[1], André O.[2], Bruno U. Marcato[2]

[1]IQSC – University of São Paulo
São Carlos, Brazil

[2]ICMC – University of São Paulo
São Carlos, Brazil.

{heitorbs, andre.to, brunoumarcato}@usp.br

Abstract. *This work investigates the feasibility of automating cortisol quantification in hair samples using Deep Learning models applied directly to raw Liquid Chromatography-Mass Spectrometry (LC-MS/MS) data. To address the labor-intensive bottleneck of manual peak integration, we evaluated an "End-to-End"Time Series Regression approach using three architectures: ROCKET, FCN, and InceptionTime. On a limited dataset ($N = 113$), ROCKET achieved the lowest Mean Squared Error (72.59), while InceptionTime demonstrated superior performance in terms of Median error (15.95) and compliance with regulatory standards (35.14% of samples within 15% error). However, statistical analysis (Friedman test, $p = 0.237$) revealed no significant difference between the models. These results indicate that while the approach is promising, the "Small Data"constraint currently limits the ability of deep models to consistently replicate analytical-grade precision, suggesting a need for larger datasets for clinical viability.*

1. Introduction

Cortisol ($11\beta, 17\alpha, 21$-trihydroxy-pregn-4-ene-3,20-dione) is a glucocorticoid steroid hormone synthesized in the adrenal cortex under the regulation of the Hypothalamic-Pituitary-Adrenal (HPA) axis. It serves as the primary biological mediator for the organism's response to physical and emotional stressors, adjusting metabolic, cardiovascular, and immunological functions [Knezevic et al. 2023]. While traditional matrices such as saliva, plasma, and urine provide acute snapshots of HPA activity, they are subject to significant circadian fluctuations, with levels peaking in the morning and declining at night. In contrast, hair analysis has emerged as a robust matrix for assessing chronic stress. As hair grows at an approximate rate of 1 cm per month, a 3 cm segment offers a retrospective calendar of systemic cortisol exposure over the preceding trimester [Stalder and Kirschbaum 2012]. Consequently, Hair Cortisol Concentration (HCC) has become a critical tool in psychoneuroendocrinology and clinical diagnostics.

The analytical gold standard for HCC quantification is Liquid Chromatography coupled to Tandem Mass Spectrometry (LC-MS/MS). In this workflow, cortisol is extracted from the keratin matrix using methanol and sonication, followed by chromatographic separation on a C18 reverse-phase column (e.g., Phenomenex Luna) using a mobile phase of water and methanol acidulated with 0.1% formic acid [de Souza Sacre 2025]. Detection is performed using a Triple Quadrupole analyzer operating in positive electrospray

ionization (ESI+) mode [Niessen 2006]. To ensure high sensitivity and selectivity in complex biological matrices, the method relies on Selected Reaction Monitoring (SRM). This involves isolating specific precursor-to-product ion transitions: *m/z* $363.09 \rightarrow 121.19$ and $363.09 \rightarrow 327.13$ for cortisol [de Souza Sacre 2025]. Crucially, quantification is performed using Isotope Dilution Mass Spectrometry, where the analyte's signal is normalized against a deuterated internal standard (Hydrocortisone-2,3,4-$^{13}C_3$, transition *m/z* $366.1 \rightarrow 124.09$) to correct for extraction efficiency, injection variability, and matrix effects [Binz et al. 2016].

However, despite the instrumental sophistication, the data processing workflow remains a significant operational bottleneck. The translation of raw ion intensity signals into concentration values requires an expert analyst to manually inspect the extracted ion chromatograms (XIC). This process involves defining baseline boundaries and verifying the integration of peak areas within the specific retention time window (approximately 6.6 minutes). This manual curation is labor-intensive, limits sample throughput, and introduces inter-operator variability, particularly when signals are affected by the moderate matrix effects often observed in hair extracts [Mirzaian et al. 2024].

Recent advancements in Deep Learning (DL) suggest that Convolutional Neural Networks (CNNs) can learn to extract features from 1D spectroscopic data, potentially automating this quantification step [Seddiki et al. 2020]. By treating the raw chromatogram as a time-series signal, DL models could theoretically map the ion intensity profile directly to the concentration, bypassing the rigid geometric rules of traditional integration algorithms [Ismail Fawaz et al. 2020]. However, analytical chemistry datasets are characteristically "Small Data"due to the high cost and complexity of sample preparation, creating a challenging environment for training deep models without overfitting.

This research aims to evaluate the feasibility of an "End-to-End"Deep Learning approach for the automated quantification of cortisol in hair, using raw LC-MS/MS data [Turova et al. 2022]. We investigate whether computational intelligence can replicate the precision of the validated analytical method without human intervention. To guide this investigation, this work addresses the following Research Questions (RQ):

- **RQ 1:** Which computational architectures are best suited for predicting analyte concentrations directly from raw LC-MS/MS signals?
- **RQ 2:** How does the performance of standard Deep Learning models, such as Fully Convolutional Networks (FCN), compare to state-of-the-art Time Series Regression architectures like InceptionTime?
- **RQ 3:** Can machine learning baselines optimized for small datasets, such as ROCKET, outperform deep neural networks in this specific chemometric domain?
- **RQ 4:** Can these automated "End-to-End"models achieve an error rate comparable to the validated manual analytical method, which typically requires a coefficient of variation (CV%) below 15%?

This research proposes methods for evaluating the feasibility of automated quantification. We explore a range of architectures to achieve this goal, comparing the statistical ROCKET method, standard FCNs, and deep InceptionTime networks. Our experimentation involves a thorough evaluation using a proprietary real dataset of hair extracts collected from post-graduate students and control groups [de Souza Sacre 2025]. This paper

extends the standard chemometric approach by applying state-of-the-art Time Series Regression (TSR) techniques to raw chromatographic data.

2. Problem Definition

Definition 2.1 (Chromatographic Signal). In the context of Liquid Chromatography-Mass Spectrometry (LC-MS/MS), a chromatographic signal refers to a time-series sequence generated by the detector. It represents the abundance (intensity) of specific ions detected over a continuous period. Mathematically, a single chromatogram channel c can be defined as a sequence of scalar intensity values $c = [i_1, i_2, ..., i_n]$, where each i_t represents the ion count at a discrete time step t, and n is the total number of scan points acquired during the acquisition window.

Definition 2.2 (Sample Instance). A sample instance is an occurrence of a chemical analysis for a specific biological specimen. Unlike a univariate time series, a sample in this domain is multivariate. It is defined as a set of parallel time series $X = \{c_1, c_2, ..., c_k\}$, where each channel c_k corresponds to a specific Selected Reaction Monitoring (SRM) transition. In this work, each hair sample instance is represented by $k = 4$ channels: the quantifier and qualifier ions for the analyte (Cortisol), and the quantifier and qualifier ions for the Deuterated Internal Standard (Cortisol-d4) [cite: 439-440]. Therefore, a single input sample is a matrix of dimensions $n \times k$ (Time $\times$ Channels).

Definition 2.3 (Quantification Task). Given a set of analyzed hair samples S, where each sample $X \in S$ consists of the raw chromatographic data described in Definition 2.2, and has a known ground-truth concentration $y \in R$ (determined by the validated analytical method involving manual integration), the problem is to learn a non-linear function $f : R^{n \times k} \rightarrow R$. This function must map the raw multichannel signal directly to the scalar concentration value, minimizing the error metric (Mean Squared Error) between the predicted concentration $\hat{y}$ and the ground truth y.

To ensure efficiency and accuracy, the data representation must account for the physical characteristics of the signal. Applications and instruments record these steps in raw data files (e.g., .mzML). To exemplify, consider a specific process instance from our dataset, Sample ID "PG_1". The raw data comprises 4 synchronized arrays of length 587 (representing approximately 10 minutes) [cite: 308-310]. The model must process these arrays, identifying the specific temporal region where the analyte elutes (retention time ≈ 6.6 min), accounting for the baseline noise and the intensity of the Internal Standard, to output the final concentration (e.g., 19.2 pg/mg).

Definition 2.4 (Small Data Constraint). A critical characteristic of this problem, which differentiates it from standard Deep Learning tasks, is the ratio of features to samples. We define the Small Data Constraint as a scenario where the dimensionality of the input X (number of time points $\times$ channels) is significantly larger than the total number of available training instances $|S_{train}|$. In this work, $|S_{total}| = 113$[cite: 1075], while the input dimensionality can exceed 2000 features (depending on the window size). This constraint necessitates the use of specific architectures capable of generalizing from limited examples, such as InceptionTime.

3. Related Work

Some works in the literature propose solutions for signal quantification and time series regression; however, the specific application of End-to-End DL for LC-MS/MS quantification remains an emerging field.

In the domain of Time Series Classification (TSC) and Regression (TSR), [Dempster et al. 2020] introduced ROCKET (Random Convolutional Kernel Transform). This approach creates a massive number of random convolutional kernels to extract features, which are then used to train a linear classifier or regressor. It is particularly noted for its computational efficiency and high accuracy on datasets with limited samples, often outperforming complex Deep Learning models. In our work, ROCKET serves as the primary Machine Learning baseline to benchmark against Deep Learning architectures.

Regarding Deep Learning for 1D signals, Fully Convolutional Networks (FCN) have been established as a strong baseline. [Wang et al. 2016] demonstrated that FCNs could effectively extract translation-invariant features from time series without manual feature engineering. However, training such networks from scratch on small datasets often leads to overfitting, a challenge addressed by [Seddiki et al. 2020], who investigated Cumulative Learning strategies to enable CNN representations for small mass spectrometry datasets.

Currently, the state-of-the-art in this domain is represented by InceptionTime [Ismail Fawaz et al. 2020]. Inspired by the Inception-v4 architecture for computer vision, this model utilizes parallel convolutional filters of varying lengths to capture both local (high-frequency) and global (low-frequency) patterns simultaneously. This is theoretically ideal for chromatography, where sharp analyte peaks (high frequency) must be distinguished from slow baseline drifts (low frequency).

Table 1 organizes the characteristics of these related works and compares them with the proposed approach.

Tabela 1. Comparison of related works and the proposed approach.

Work	Architecture	Input Data	Domain	Original Task
[Niessen 2006]	Standard Analytical	Integrated Peak Area	Chromatography	Quantification
[Wang et al. 2016]	FCN	Raw Time Series	General (UCR)	Classification
[Dempster et al. 2020]	ROCKET	Raw Time Series	General (UCR)	Classification
[Ismail Fawaz et al. 2020]	InceptionTime	Raw Time Series	General (UCR)	Classification
[Seddiki et al. 2020]	1D-CNN	Raw MS Spectra	Mass Spectrometry	Classification
This work	**InceptionTime/ROCKET**	**Raw MS Spectra**	**Mass Spectrometry**	**Quantification**

4. Methodology

Figure 1 outlines the methodological workflow adopted in this study. A detailed description of each component is provided below.

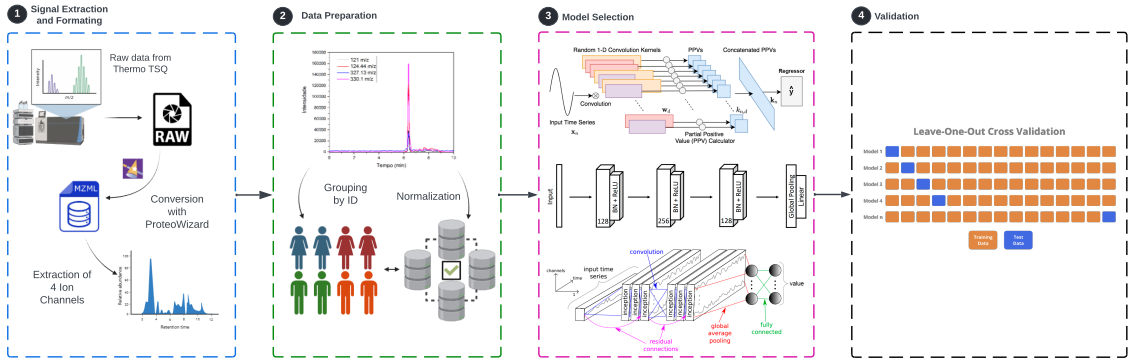

Figura 1. Methodological workflow

4.1. Signal extraction and formating

The raw spectral data were acquired using a Thermo Scientific TSQ Quantum Access triple quadrupole mass spectrometer, resulting in proprietary binary files (`.raw` format). To enable computational processing and ensure interoperability with open-source data mining libraries, these files were converted to the open standard `.mzML` (Mass Spectrometry Markup Language) format. This conversion was performed using the *ProteoWizard msconvert* utility, utilizing the "Peak Picking"filter (vendor algorithm) to centroid the data and "zlib"compression to optimize storage. The resulting `.mzML` files were then parsed using the *PyOpenMS* library to extract the intensity arrays corresponding to the four monitored SRM transitions, enabling their transformation into the tensor format required for the neural networks.

From the now formatted data, we extracted the four Selected Reaction Monitoring (SRM) chromatograms corresponding to the quantifier and qualifier ions for both the analyte (Cortisol) and the Internal Standard (Cortisol-d4). Unlike traditional approaches that reduce these signals to a single scalar value (peak area), we retained the complete temporal resolution of the chromatographic run. Each sample was thus represented as a multivariate time series tensor of shape (T, C), where T represents the number of scan points acquired over the 10-minute chromatographic run (approximately 587 time points) and $C = 4$ represents the ion channels. This "End-to-End"representation preserves all morphological features of the peak, including baseline drift, peak symmetry, and potential co-eluting interferences, allowing the model to learn feature extraction directly from the raw signal context.

4.2. Data pre-processing

To transform the raw instrumental data into a format suitable for Time Series Regression architectures, a rigorous pre-processing pipeline was established. The raw data files, originally in `.raw` format, underwent three primary stages of transformation to ensure data quality and model stability.

4.2.1. Channel-wise Z-score normalization

Neural networks are highly sensitive to the scale of input data, particularly in mass spectrometry where ion intensities can span several orders of magnitude (typically from 10^3 to

10^7 counts). Furthermore, the intensity difference between the abundant Internal Standard and the trace-level Cortisol can be significant. To address this, we applied Channel-wise Z-score normalization. Instead of normalizing the entire sample by a single scalar, each of the 4 channels was normalized independently. For a given sample x, the intensity I at time t for channel c was transformed according to:

$$I'_{t,c} = \frac{I_{t,c} - \mu_c}{\sigma_c} \tag{1}$$

where μ_c and σ_c are the mean and standard deviation of the intensities for that specific channel c, calculated across the entire training dataset.

4.3. Model Selection

Three time series regression models were selected for the cortisol quantification proposed in this study: a baseline model (Rocket), a standard neural network (FCN), and a state-of-the-art model (InceptionTime). A brief overview of their underlying mechanisms is provided in Section 3.

The Aeon library [Ermshaus et al. 2023], which provides implementations for all the aforementioned models, was employed in this work. While various hyperparameters allow for model fine-tuning, to ensure experimental reproducibility, the specific configurations used in this study are detailed in Table 2.

Tabela 2. Hyperparameters and configurations for the regression models.

Model	Parameter	Value
Rocket	Number of Kernels	10,000
	Regression Method	RidgeCV
FCN	Network Depth	3 Layers
	Filters per Layer	{128, 256, 128}
	Kernel Size	{8, 5, 3}
	Activation Function	ReLU
	Epochs	100
	Batch Size	8
	Loss	Mean Squared Error
	Optimization	Adam
InceptionTime	Ensemble Size	5
	Inception Modules	6 Modules
	Filters per Module	32
	Conv. Layers per Module	3
	Kernel Size	40
	Bottleneck Size	32
	Activation Function	ReLU
	Epochs	100
	Batch Size	8
	Loss	Mean Squared Error
	Optimization	Adam

Furthermore, to prevent overfitting and unnecessary computation during training stagnation, an Early Stopping callback was employed for both FCN and InceptionTime. This mechanism monitored the loss metric, halting execution if no improvement was observed over 20 consecutive epochs (patience) and automatically restoring the model weights that yielded the best performance.

4.4. Data Partitioning and Validation

A critical challenge in this study was the limited sample size ($N = 113$) and the presence of biological replicates (duplicate injections from the same subject). Standard random splitting would inevitably lead to *data leakage*, where a subject's duplicate appears in the test set while their primary sample is included in the training set, artificially inflating performance metrics. To ensure a rigorous and unbiased evaluation, we employed **Leave-One-Out Cross-Validation (LOOCV)** with grouping by Subject ID. In this scheme, each iteration holds out all chromatographic replicates belonging to a single subject as the test partition, while all remaining subjects compose the training set. This strict subject-level separation prevents the model from memorizing sample-specific artifacts and forces it to generalize to entirely unseen individuals, yielding a more realistic estimate of predictive performance in a clinical context.

Finally, the source code and datasets employed in this work are publicly available on GitHub[1] to facilitate the reproduction of this study.

5. Experiments and results

The experiments were conducted in accordance with the methodology previously outlined. Given the regression nature of the problem, the Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Median were selected as performance metrics. The results obtained, along with other relevant indicators, are presented in Table 3.

Tabela 3. Performance comparison of the regression models.

Model	MSE	SD (MSE)	RMSE	Median	$y < 0.15\bar{x}$	$y < 0.15\bar{x}$ (%)
ROCKET	**72.59**	**108.66**	**8.52**	28.19	41	27.70
FCN	115.91	260.22	10.77	29.81	48	32.43
InceptionTime	84.81	162.62	9.21	**15.95**	**52**	**35.14**

The data indicates high mean errors and standard deviations across all models. While the error magnitude must be contextualized against the ground truth data mean of 16.24, the RMSE remains significant; even the best-performing model in this regard, ROCKET, achieved a value of 8.52. Furthermore, the high Standard Deviation of the MSE across all models demonstrates a substantial variance in the predictions. This variability is visually depicted in Figure 2, which presents boxplots for all evaluated models. To ensure a clearer visualization of the distribution, outliers have been omitted from the plot.

[1] https://github.com/andryll/Cortisol-Estimator-with-DL

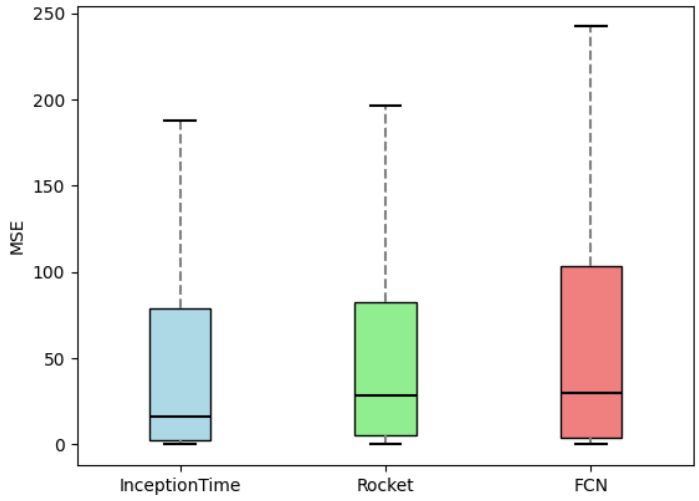

Figura 2. Error distribution boxplots for each model (outliers omitted).

However, the Median metric offers a crucial alternative perspective. Although ROCKET achieved the lowest overall MSE and SD, InceptionTime presented a considerably lower median value—nearly half that of the other models. This discrepancy reveals that while InceptionTime yields superior accuracy for at least half of the samples (as indicated by the median), its performance on the remaining data points likely involves severe outliers that skew the mean-based metrics (MSE/RMSE) upward.

A further critical perspective involves the validation criteria for bioanalytical methods established by the Brazilian Health Regulatory Agency (ANVISA) through Resolution RDC No. 27/2012 [ANVISA 2012]. According to Art. 30 of this regulation, the accuracy of a quantitative method is considered acceptable if the relative error remains within $\pm 15\%$ of the nominal value. In this context, Table 3 reports—in columns $y < 0.15\bar{x}$ and $y < 0.15\bar{x}$ (%)—the number and percentage of samples falling within this regulatory threshold. Notably, despite yielding a higher mean error compared to ROCKET, InceptionTime achieved the highest number of estimations compliant with ANVISA's criteria. However, the resulting compliance rate (35.14%) indicates that the current model performance does not yet fully satisfy the strict bioanalytical validation requirements.

Finally, a Friedman test was performed to analyze the statistical significance of the performance differences among the models. The test yielded a Chi-Square value of 2.716 and a p-value of 0.237. Consequently, it is concluded that there is no statistically significant difference between the models.

6. Conclusion and Future Works

Based on the presented results, it can be concluded that the automatic estimation of cortisol levels using neural networks represents a promising avenue of research. However, given that current performance metrics fell short of initial expectations, further experimentation is required.

A significant challenge encountered in this study was the scarcity of data relative to the complexity of the task, which likely contributed to the high variance observed in the models' predictions. It is hypothesized that incorporating a larger dataset would yield more consistent results, potentially enhancing the overall performance across all evaluated models.

6.1. Future Works

Future work should prioritize the development of mass spectrometry data synthesis methods that accurately reflect the characteristics of the real-world data used in this study. This approach would facilitate the generation of the large-scale datasets required for effective training, thereby improving model performance. Given that the acquisition of real biological data is a complex and labor-intensive process, the utilization of synthetic data represents a viable solution to the data scarcity problem.

Another avenue for investigation involves the evaluation of alternative machine learning architectures for cortisol level estimation. Initial candidates include neural networks such as ResNet and LITETime. Furthermore, non-neural approaches, such as HIVE-COTE, also warrant investigation, although their deployment must be carefully weighed against their significantly higher computational costs.

Referências

ANVISA (2012). Resolution RDC no. 27, of may 17, 2012. Official Gazette of the Federal Executive. Establishes minimum requirements for the validation of bioanalytical methods employed in studies for registration and post-registration of medicines. Brasília, Brazil.

Binz, T. M., Braun, U., Baumgartner, M. R., and Kraemer, T. (2016). Development of an lc–ms/ms method for the determination of endogenous cortisol in hair using 13c3-labeled cortisol as surrogate analyte. *Journal of Chromatography B*, 1033-1034:65–72.

de Souza Sacre, M. E. (2025). Desenvolvimento de método bioanalítico para avaliação de cortisol capilar como indicador de estresse acadêmico. Dissertação de mestrado, Instituto de Química de São Carlos, Universidade de São Paulo (USP), São Carlos, SP.

Dempster, A., Petitjean, F., and Webb, G. I. (2020). Rocket: exceptionally fast and accurate time series classification using random convolutional kernels. *Data Mining and Knowledge Discovery*, 34(5):1454–1495.

Ermshaus, A., Schäfer, P., Bagnall, A., Guyet, T., Ifrim, G., Lemaire, V., Leser, U., Leverger, C., and Malinowski, S. (2023). Human activity segmentation challenge @ ecml/pkdd'23. In Ifrim, G., Tavenard, R., Bagnall, A., Schaefer, P., Malinowski, S., Guyet, T., and Lemaire, V., editors, *Advanced Analytics and Learning on Temporal Data*, pages 3–13, Cham. Springer Nature Switzerland.

Ismail Fawaz, H., Lucas, B., Forestier, G., Pelletier, C., Schmidt, D. F., Weber, J., Webb, G. I., Idoumghar, L., Muller, P.-A., and Petitjean, F. (2020). Inceptiontime: Finding alexnet for time series classification. *Data Mining and Knowledge Discovery*, 34(6):1936–1962.

Knezevic, E., Nenic, K., Milanovic, V., and Knezevic, N. N. (2023). The role of cortisol in chronic stress, neurodegenerative diseases, and psychological disorders. *Cells*, 12(23).

Mirzaian, M., van Zundert, S. K., Schilleman, W. F., Mohseni, M., Kuckuck, S., van Rossum, E. F., van Schaik, R. H., and van den Berg, S. A. (2024). Determination of cortisone and cortisol in human scalp hair using an improved lc-ms/ms-based method. *Clinical Chemistry and Laboratory Medicine (CCLM)*, 62(1):118–127.

Niessen, W. M. A. (2006). *Liquid Chromatography–Mass Spectrometry*. CRC Press, Taylor & Francis Group, Boca Raton, FL, 3rd edition.

Seddiki, K., Saudemont, P., Precioso, F., et al. (2020). Cumulative learning enables convolutional neural network representations for small mass spectrometry data classification. *Nature Communications*, 11(1):5595.

Stalder, T. and Kirschbaum, C. (2012). Analysis of cortisol in hair – state of the art and future directions. *Brain, Behavior, and Immunity*, 26(7):1019–1029.

Turova, P., Stavrianidi, A., Svekolkin, V., Lyskov, D., Podolskiy, I., Rodin, I., Shpigun, O., and Buryak, A. (2022). Analysis of primary liquid chromatography mass spectrometry data by neural networks for plant samples classification. *Metabolites*, 12(10).

Wang, Z., Yan, W., and Oates, T. (2016). Time series classification from scratch with deep neural networks: A strong baseline.