Proceedings Paper:
Wang, B., Wu, Y., Vaci, N. orcid.org/0000-0002-8094-0902 et al. (3 more authors) (2021) Modelling paralinguistic properties in conversational speech to detect bipolar disorder and borderline personality disorder. In: ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing, 06-11 Jun 2021, Toronto, Ontario, Canada. Institute of Electrical and Electronics Engineers , pp. 7243-7247. ISBN 9781728176062

https://doi.org/10.1109/icassp39728.2021.9413891

MODELLING PARALINGUISTIC PROPERTIES IN CONVERSATIONAL SPEECH TO DETECT BIPOLAR DISORDER AND BORDERLINE PERSONALITY DISORDER

Bo Wang[1,3], *Yue Wu*[2,3], *Nemanja Vaci*[4], *Maria Liakata*[3,5], *Terry Lyons*[2,3], *Kate E A Saunders*[1]

[1]Department of Psychiatry and [2]Mathematical Institute, University of Oxford, UK
[3]The Alan Turing Institute, London, UK
[4]Department of Psychology, University of Sheffield, UK
[5]School of Electronic Engineering and Computer Science, Queen Mary University of London, UK
bo.wang@psych.ox.ac.uk

ABSTRACT

Bipolar disorder (BD) and borderline personality disorder (BPD) are two chronic mental health conditions that clinicians find challenging to distinguish based on clinical interviews, due to their overlapping symptoms. In this work, we investigate the automatic detection of these two conditions by modelling both verbal and non-verbal cues in a set of interviews. We propose a new approach of modelling short-term features with visibility-signature transform, and compare it with widely used high-level statistical functions. We demonstrate the superior performance of our proposed signature-based model. Furthermore, we show the role of different sets of features in characterising BD and BPD.

Index Terms— bipolar disorder, borderline personality disorder, speech analysis, paralinguistic modelling, path signature

1. INTRODUCTION

Bipolar disorder (BD) is a mood disorder characterised by extreme mood swings between manic highs and depressive lows that can last from days to weeks each. Borderline personality disorder (BPD) is a type of personality disorder marked by a long-term pattern of varying moods, self-image and behaviour. They both can seriously affect the patients' ability to function in work and social activities [1, 2, 3]. They also can co-occur in 10%-20% of the cases and since the symptomatology of these two disorders is very similar, differentiating between them poses a diagnostic challenge for the clinicians [4, 5]. However, accurate diagnosis is crucial, as the most effective treatment for BD being pharmacological while for BPD psychotherapeutic. Standard diagnostic assessment for BD and BPD rely primarily on clinical interviews where the patients have to describe their accounts of symptoms. As a result it can be potentially influenced by patients' retrospective recall biases and cognitive limitations [6].

In recent years a number of machine learning based studies have investigated the use of acoustic features from speech for automated assessment across several psychiatric disorders showing diagnostic potentials [6, 7, 8]. Existing work on bipolar disorder showed mood episodes affecting patients' speech and in turn acoustic and dialogue features extracted from speech can be used in detecting mood states [9, 10, 11]. In this work, we aim to model both verbal and non-verbal cues in non-clinical interviews for the distinction between BD and BPD, which remains understudied.

One key step of dialogue-based automated assessment system is modelling a series of conversational utterances and aggregate features from each segment to a fixed-size representation for down-stream classification or regression. Traditionally, a very popular approach has been to apply a set of high-level statistical functions (e.g. mean, max, variance, linear regression coefficients, etc.) for feature aggregation [9, 10, 11, 12], where the role of these aggregation functions is to describe the global characteristics of given spoken conversation. However, the conversational dynamics is not effectively modelled during this process and important sequential information may be ignored as a result. Recurrent neural networks (RNN) are designed to explicitly model sequential data, however, they are prone to overfit on small datasets. Signature transform, initially introduced in stochastic analysis, is a non-parametric approach of encoding sequential data and capturing the order information in the data. It has been proven effective in a range of machine learning tasks involving sequential modelling [13, 14, 15]. A recent study [16] proposed summarising linguistic and turn-taking behaviour features from recorded interviews using signature. However, minimal description of its working was provided, and no comparative evaluation was given.

The contributions of this work are as follows: (1) We study three types of acoustic features for the classification of BD and BPD patients; (2), We propose using visibility transform to enhance the signature representation of conversational speech; (3), We evaluate and compare between high-level statistical functions (HSF) and signature transform (SIG) based models, and show the superiority of SIG.

2. AMOSS-I DATASET

The Automated Monitoring of Symptoms Severity Interview (AMoSS-I) dataset [16] contains 50 participants, who were interviewed and transcribed to gather qualitative feedback of the original AMoSS study. Among these participants, 21 had a BD diagnosis, 17 had been diagnosed with BPD and 12 were healthy controls. Each participant was interviewed only once.

Among the 50 one-on-one qualitative interviews in AMoSS-I, 32 were recorded in the meeting room while the remaining 18 were phone interviews. As summarised in Table 1, the phone interviews are more likely to be shorter than the ones conducted in the meeting room. We also take the difference of the peak and trough values for Root-mean-square (RMS) level measured over a 10 ms window, showing the difference of loudness within each audio recording. Comparing to the meeting room interviews, on average the phone interviews are shown to have significant larger difference of loudness within each one, mainly due to the way of recording. Additionally, we find the noise level in the room interviews to be higher as seen in their much lower signal-to-noise ratio of 21.16 dB, computed by Waveform Amplitude Distribution Analysis (WADA-SNR) [17].

Table 1. Differences in sample size and acoustics between the room and phone interviews. N is the number of the interviews. Audio length in minutes summarised in the form of the median +/- the interquartile range. The percentages of clipped samples are averaged over each set of interviews. RMS diff is the difference between the peak and trough values for RMS level measured over a 10 ms window, in dBFS. Signal-to-noise ratio (SNR) measures the amount of non-speech in a speech signal in decibels (dB).

Env	N	Length	%Clipped	RMS diff	SNR
Room	32	23 ± 12	10.28%	64.91	21.16
Phone	18	19 ± 10	3.48%	83.10	95.30
Both	50	22 ± 11	7.83%	71.46	47.85

2.1. Data Preprocessing

The two different recording environments also resulted in different levels of clipping. As shown in Table 1, clipping occurs much more often in the meeting room interview recordings than the phone interviews, with an average of 10.28% of the audio signals being clipped from their maximum range. We have reduced the level of clipping by extrapolating the clipped parts of the audio using an open-source digital audio editor *Audacity*[1]. In order to alleviate the effect of loudness difference shown in Table 1, we scale the audio signal for each speaker turn separately, and make sure each turn is in the range of -1 and 1. We apply a domain-adversarial neural network based voice activity detection model [18] for intra-speaker-turn segmentation. We also exclude speaker turns that are shorter than 2 seconds to extract robust acoustic features.

3. FEATURE EXTRACTION

Instead of using more complex features such as mel frequency cepstral coefficients (MFCCs) or high-dimensional embeddings from pretrained speech encoders, we identify a set of acoustic and (non-verbal) turn-taking behaviour related features for their interpretability.

Prosodic features: During these dyadic interviews, participants often exhibited changes in prosodic variables such as the tone and intonation of their speech when they are recollecting their experience especially from the intensive week of the study. We compute Legendre polynomial coefficients for *fundamental frequency* (F_0) and *energy contours*. Together with *segment duration* they form 13 dynamic features to capture the prosodic variations. Prosodic features have been proven effective in identifying mood states in BD [6, 9].

Rhythm features: We also extract 7 rhythm features from each speaker-turn using algorithm proposed by Tilsen and Arvaniti [19] that is based on empirical mode decomposition of the vocalic energy amplitude envelope. The envelope is decomposed into two intrinsic mode functions (IMF). They show the IMF-based rhythm features can capture information about periodicities that likely correspond to different linguistic constructs, and thus are useful for examining rhythmicity in speech. Previous studies also used them for mood state detection [10, 11].

Phonation features: Voice quality has an important role in signalling paralinguistic information [20], and previous study showed significant difference in the speaker's voice quality when comparing people who suffer from psychological disorder to healthy controls [21]. We compute 7 phonation-based dynamic features from sustained vowels and continuous speech utterances, including *jitter*, *shimmer*, *amplitude perturbation quotient* and *pitch perturbation quotient*.

Dialogue features: To model high-level interactive patterns in the dialogue, we extract the set of 13 turn-taking behaviour related features following the work in [11, 16], including *relative floor control*, *turn hold offset*, *number of consecutive turns*, *turn switch offset*, *speech overlaps* and *number of words per second*, per turn.

All acoustic and dialogue features are Z-normalised either: 1) using the mean and standard deviation of each interview respectively, denoted as "Person", or 2) using the mean and standard deviation of all training samples, denoted as "Global". Previous studies [10, 22] suggest normalisation by "Person" focuses on each individual baseline and can reduce the potential extraneous effect, for example, caused by different recording environment.

[1] https://www.audacityteam.org/

4. FEATURE AGGREGATION

The paralinguistic properties drawn from a patient's response in an interview, represented by the features described in Section 3, can be indicative of the state of the patient at the time of speaking. However, the state of the patient also evolves while being affected by the highly dynamic nature of a dyadic conversation. The common approach of applying high-level statistical functions (HSFs) does not effectively capture this dynamics. In this section we describe the method of signature transform (SIG), which we use to aggregate frame-level acoustic or turn-level dialogue features and encode nonlinear time-dependent interactions in the feature set, $S(X)$.

A sequentially ordered data stream may be thought of as a discretisation of a path of finite length $X : [a,b] \to \mathbb{R}^d$, where $a \leq b$ and $d \in \mathbb{N}$. For example, a speech sequence represented by its phonation-based dynamic features can be thought of as a path of $d = 7$. *Signature transform*, also known as *path signature*, describes a graded sequence of statistics characterising the underlying path [23, 24], and thus provides an effective feature set for capturing its total ordering (i.e., incremental effect) while ignoring the positional effect. More specifically, consider a d-dimensional path X over the time interval $[a,b]$, the signature $S(X)$ of this path is the infinite collection of statistics[2]:

$$S(X)_{a,b} = \left(\left\{ S(X)^{i_1}_{a,b} \right\}_{1\leq i_1\leq d}, \left\{ S(X)^{i_1,i_2}_{a,b} \right\}_{1\leq i_1,i_2\leq d}, \dots \right)$$

where each term is a n-fold iterated integral of x with multi-index $i_1, \dots, i_k$:

$$S(X)^{i_1,\dots,i_k}_{a,b} = \int\cdots\int_{\substack{t_1<\cdots<t_k \\ t_1,\dots,t_k\in[a,b]}} dX^{i_1}_{t_1} \otimes \cdots \otimes dX^{i_k}_{t_k}$$

where $k \in \mathbb{N}$. $S(X)^{i_1,\dots,i_k}_{a,b}$ is termed as the kth level of the signature. In practice we truncate the signature to order n to have a finite dimensional representation. Note the positional information in the data stream can be informative in many applications. For example, it's useful to know the beginning and end of each speaker turn when modelling the conversation. The visibility transform, initially introduced in [26], is able to embed the effect of the absolute positions of the input sequence into signature transform, which is otherwise not included in the signature as it is translation-invariant map[3].

Consider a d-dimensional data sequence of length n, i.e., $\mathrm{x} = (x_1, \dots, x_n)$. The visibility transform in prior to computing signature, adds two time steps and a binary coordinate to the input sequence x that is equal to 1 until the second-to-last time step:

$$\phi\left(\mathrm{x}\right) = \left(\mathrm{ap}_1(x_1), \dots, \mathrm{ap}_1(x_{n-1}), \mathrm{ap}_1(x_n), \mathrm{ap}_0(x_n), \mathbf{0}\right),$$

where $\mathrm{ap}_c : \mathbb{R}^d \to \mathbb{R}^{d+1}$ is an operator expanding a d-dimensional vector to $d+1$ dimensions by appending scalar c at the end, and $\mathbf{0} \in \mathbb{R}^{d+1}$. The resulting sequence $\phi\left(\mathrm{x}\right)$ is $d+1$-dimensional and of length $n+2$. We apply visibility transform to our frame-level acoustic features per speaker-turn before feature aggregation. This way the start and end positions of each turn can be embedded in the interview-level signature representation. We name this approach as VT-SIG.

5. EXPERIMENTS AND ANALYSIS

Following previous work we choose nested leave-one-subject-out as the evaluation scheme, and logistic regression with L2 regularisation for classification. For each fold, we first apply VT-SIG to each feature type, and keep only the first three levels of the path signature[4]. We conduct feature selection on signature-transformed interview-level features through computing Pearson Correlation Coefficients (PCC) with the IPDE scores[5] on the training data and retain the features with p-values less than 0.001. This results in a small number of features. The selected features are then fed to the classifier for 3 separate binary tasks: (1), BD vs. healthy controls, (2), BPD vs. healthy controls, and (3), BD vs. BPD patients.

5.1. Analysis of the selected features

Five most significant and commonly selected features from each task are briefly summarised in Table 2 as examples. Each interview-level feature (a signature term) is represented as a linear combination of the original frame-level or turn-level features. We see almost all of the selected features are volume integrals, i.e., they are triple integrals of three acoustic/dialogue features. We notice the importance of the binary coordinate (i.e., *c* in Table 2) from visibility transform, to the task of *H vs. BD*. For example, *(c, apq, logE)* represents the nonlinear effect between *apq* and *logE* only at the last frame of each turn of speech[6] while its feature values before the last frame are shown to be unimportant. *(IMF1_m, c, c)* and *(c, IMF1_m)*, which represent the linear incremental effect of *IMF1_m* and the linear effect of *IMF1_m* at the last frame (of each turn) respectively, are also among the selected features.

While most of the important features for *H vs. BD* represent either purely incremental or (last-)positional effects, the selected signature terms for the other two tasks are of nonlinear mixed effects from both aspects. Overall, we notice the importance of the phonation features in two tasks involving BD and in particular, *BD vs BPD*. Rhythm features, which

[2] We refer the readers to [15] Appendix A for a formal definition of signature transform, and [25] for a primer on its use in machine learning.

[3] The mathematical definition of the visibility transform can be found in [27].

[4] We use iisignature Python library, `https://pypi.org/project/iisignature/`.

[5] The International Personality Disorder Examination (IPDE) [28] is a semi-structured clinical interview designed to assess major categories of personality disorders.

[6] If *c* is indexed first in the signature term, then this term only captures the effect from the coordinates of the following indices at the last time step. This is proved in Theorem 5 of [27].

have previously used for mood state detection, are selected in distinguishing between the BD/BPD patients and healthy controls. For each acoustic/dialogue feature, we also analyse its potential interaction effect with the recording environment, who the interviewer is and the gender of the participant, which we treat as control variables, by fitting a linear model with diagnosis as response. We find the 'recording environment' variable significantly changed the effect of *(apq, c, logE)* ($p < 0.05$). As a result we remove it from the final feature set for classification, this way we only model effects that are not systematically variable across situational factors.

Table 2. Top-5 most significant and commonly selected features during LOOCV. Features belong to the dialogue category are colored in red; rhythm features are in green and phonation features are in blue.[7]

H vs BD	H vs BPD	BD vs BPD
(c, apq, logE)	*(SPBr, IMF12, IMF2_m)*	*(DF0, Jitter, logE)*
(c, logE, apq)	*(CNTR, IMF12, IMF2_m)*	*(logE, DDF0, Jitter)*
(apq, c, logE)	*(TL, n_OVL, TSO)*	*(logE, DF0, Jitter)*
(c, IMF1_m)	*(n_OVL, RFC_t, SP_m)*	*(logE, Jitter, DDF0)*
(IMF1_m, c, c)	*(n_OVL, RFC_t, n_LP)*	*(ppq, logE, DF0)*

5.2. Results and Discussion

We first summarise the classification results obtained by using visibility transform enhanced signature (VT-SIG) in Table 3. With the (late) fusion of acoustic and dialogue features extracted from participants' speech only, we obtain a AUROC of 0.717 in H/BD, 0.841 in H/BPD and 0.716 in BD/BPD. The classification performance drop sharply when we switch to interviewers' speech instead, due to the non-clinical nature of the interviews. Modelling the interviews as a sequence of turns from both speakers (named "Both") also result in worse performance than learning from only the participants.

Table 3. Classification results for three binary tasks: H vs. BD, H vs. BPD and BD vs. BDP, using logistic regression. Results shown are macro-averaged F_1 and AUROC scores across all interviews. p-value used for feature selection: '*'$<$ 0.005, or else $<$ 0.001 is used.

	H/BD		**H/BPD**		**BD/BPD**	
Subject	**F_1**	**AUC**	**F_1**	**AUC**	**F_1**	**AUC**
Participant	**0.738**	**0.738**	**0.827**	**0.841**	**0.710**	**0.716**
Interviewer	0.581	0.583	0.102*	0.100*	0.552*	0.556*
Both	0.477	0.488	0.512*	0.515*	0.683	0.686

We also compare results obtained by using different feature normalisation and aggregation methods described in Section 3 and 4. First, we notice VT-SIG in general has the better performance and is more reliable across all three tasks. As for HSF though we have to increase the p-value feature selection threshold from 0.001 to 0.005 or even 0.01 to have any feature for H/BD and H/BPD, it still obtains relatively poor performance. Secondly, it is also shown normalising features per interview ("Person") has led to increased performance for signature-based models with and without the use of visibility transform (namely, VT-SIG and SIG). Different feature normalisation methods have not made any significant impact on the overall performance of the HSF-based models.

Table 4. Performance comparison among different feature aggregation (Aggr)[8] and normalisation methods (Norm). Results shown are average **AUROC**s across all interviews. p-value codes: '**'<0.001; '*'<0.005; '†'<0.01.

Aggr	**Norm**	**H/BD**	**H/BPD**	**BD/BPD**
HSF	None	0.381†	0.544†	0.681**
HSF	Global	0.405†	0.515†	0.686**
HSF	Person	0.575*	0.556*	0.543**
SIG	None	0.554*	0.699**	0.644**
SIG	Global	0.435*	0.699**	0.662**
SIG	Person	0.661*	0.686**	0.716**
VT-SIG	None	0.608**	0.728**	0.710**
VT-SIG	Global	0.554**	0.811**	0.633**
VT-SIG	Person	**0.738****	**0.841****	**0.716****

6. CONCLUSIONS AND FUTURE WORK

In this paper, we demonstrate the potential of using speech from non-clinical interviews for detecting BD and BPD. Modelling short-term features and generating final representation is key for any machine learning based mental health assessment model. We propose the use of visibility-signature transform that embeds sequential ordering for feature aggregation. We show the better performance obtained by the proposed approach comparing with widely used high-level statistical functions. For future work, we plan for new data collection with more participants and multiple interviews per subject, from two different locations, which will allow for longitudinal studies and cross-site validation.

7. ACKNOWLEDGEMENTS

This work was supported by the MRC Mental Health Data Pathfinder award to the University of Oxford [MC_PC_17215], by the NIHR Oxford Health Biomedical Research Centre and by the The Alan Turing Institute under the EPSRC grant EP/N510129/1. We would also like to thank John Gideon and Fasih Haider for sharing code and thoughts.

[7] *c*: binary coordinate added during visibility transform; *apq*: amplitude perturbation quotient, *logE*: logaritmic energy; *DF0/DDF0*: first/second derivative of the fundamental Frequency; *IMF1_m*: mean within-utterance instantaneous freq. of IMF1; *IMF12*: ratio between IMF2 and IMF1; *SPBr*: ratio between power in envelope spectrum bands (1/3.5/10 Hz); *CNTR*: envelope spectrum centroid computed over 1-10 Hz band; *TL*: duration of each speaker turn; *n_OVL*: number of speech overlaps; *TSO*: latency between speaker turn transitions; *RFC_t*: relative floor control (time); *SP_m*: average length of short pauses; *n_LP*: number of long pauses (>500 ms).

[8] For reasonable comparison we adopt a wide range of HSFs previously used in [10].

8. REFERENCES

[1] W. Coryell, W. Scheftner, M. Keller, J. Endicott, J. Maser, and G. L. Klerman, “The enduring psychosocial consequences of mania and depression,” *The American journal of psychiatry*, 1993.

[2] J. F. Goldberg, M. Harrow, and L. S. Grossman, “Course and outcome in bipolar affective disorder: a longitudinal follow-up study,” *The American journal of psychiatry*, 1995.

[3] M. Zimmerman, W. Ellison, T. A. Morgan, D. Young, I. Chelminski, and K. Dalrymple, “Psychosocial morbidity associated with bipolar disorder and borderline personality disorder in psychiatric out-patients: comparative study,” *The British Journal of Psychiatry*, 2015.

[4] M. Zimmerman and T. A. Morgan, “The relationship between borderline personality disorder and bipolar disorder,” *Dialogues in clinical neuroscience*, 2013.

[5] M. Zimmerman, J. H. Martinez, T. A. Morgan, D. Young, I. Chelminski, and K. Dalrymple, “Distinguishing bipolar ii depression from major depressive disorder with comorbid borderline personality disorder: demographic, clinical, and family history differences,” *The Journal of clinical psychiatry*, 2013.

[6] D. M. Low, K. H. Bentley, and S. S. Ghosh, “Automated assessment of psychiatric disorders using speech: A systematic review,” *Laryngoscope Investigative Otolaryngology*, vol. 5, no. 1, pp. 96–116, 2020.

[7] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, “A review of depression and suicide risk assessment using speech analysis,” *Speech Communication*, vol. 71, pp. 10–49, 2015.

[8] A. Parola, A. Simonsen, V. Bliksted, and R. Fusaroli, “Voice patterns in schizophrenia: A systematic review and bayesian meta-analysis,” *Schizophrenia Research*, vol. 216, pp. 24–40, 2020.

[9] M. Faurholt-Jepsen, J. Busk, M. Frost, M. Vinberg, E. M. Christensen, O. Winther, J. E. Bardram, and L. V. Kessing, “Voice analysis as an objective state marker in bipolar disorder,” *Translational psychiatry*, 2016.

[10] J. Gideon, E. M. Provost, and M. McInnis, “Mood state prediction from speech of varying acoustic quality for individuals with bipolar disorder,” in *ICASSP*, 2016.

[11] Z. Aldeneh, M. Jaiswal, M. Picheny, M. McInnis, and E. M. Provost, “Identifying mood episodes using dialogue features from clinical interviews,” in *Interspeech*, 2019.

[12] K. Matton, M. G. McInnis, and E. M. Provost, “Into the wild: Transitioning from recognizing mood in clinical interactions to personal conversations for individuals with bipolar disorder.,” in *Interspeech*, 2019.

[13] I. P. Arribas, K. Saunders, G. Goodwin, and T. Lyons, “A signature-based machine learning model for bipolar disorder and borderline personality disorder,” *Translational Psychiatry*, p. 274, 2018.

[14] B. Wang, M. Liakata, H. Ni, T. Lyons, A. J. Nevado-Holgado, and K. Saunders, “A path signature approach for speech emotion recognition,” in *Interspeech*, 2019.

[15] P. Kidger, P. Bonnier, I. P. Arribas, C. Salvi, and T. Lyons, “Deep signature transforms,” in *Advances in Neural Information Processing Systems*, 2019.

[16] B. Wang, Y. Wu, N. Taylor, T. Lyons, M. Liakata, A. J. Nevado-Holgado, and K. E. Saunders, “Learning to detect bipolar disorder and borderline personality disorder with language and speech in non-clinical interviews,” in *Interspeech*, 2020.

[17] C. Kim and R. M. Stern, “Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis,” in *Interspeech*, 2008.

[18] M. Lavechin, M.-P. Gill, R. Bousbib, H. Bredin, and L. P. Garcia-Perera, “End-to-end domain-adversarial voice activity detection,” in *Interspeech*, 2020.

[19] S. Tilsen and A. Arvaniti, “Speech rhythm analysis with decomposition of the amplitude envelope: characterizing rhythmic patterns within and across languages,” *The Journal of the Acoustical Society of America*, 2013.

[20] N. Campbell and P. Mokhtari, “Voice quality: the 4th prosodic dimension,” in *15th ICPhS*, 2003.

[21] S. Scherer, G. Stratou, J. Gratch, and L.-P. Morency, “Investigating voice quality as a speaker-independent indicator of depression and ptsd.,” in *Interspeech*, 2013.

[22] N. Cummins, J. Epps, M. Breakspear, and R. Goecke, “An investigation of depressed speech detection: Features and normalization,” in *Interspeech*, 2011.

[23] K.-T. Chen, “Integration of paths-a faithful representation of paths by non-commutative formal power series,” *Transactions of the American Mathematical Society*, vol. 89, no. 2, pp. 395–407, 1958.

[24] T. J. Lyons, M. Caruana, and T. Lévy, *Differential equations driven by rough paths*, Springer, 2007.

[25] I. Chevyrev and A. Kormilitzin, “A primer on the signature method in machine learning,” *arXiv preprint arXiv:1603.03788*, 2016.

[26] W. Yang, T. Lyons, H. Ni, C. Schmid, L. Jin, and J. Chang, “Leveraging the path signature for skeleton-based human action recognition,” *arXiv preprint arXiv:1707.03993*, 2017.

[27] Y. Wu, H. Ni, T. J. Lyons, and R. L. Hudson, “Signature features with the visibility transformation,” in *25th International Conference on Pattern Recognition*, 2020.

[28] A. W. Loranger, A. Janca, and N. Sartorius, *Assessment and diagnosis of personality disorders: The ICD-10 international personality disorder examination (IPDE)*, Cambridge University Press, 1997.