Real-Time Sociometrics from Audio-Visual Features for Two-Person Dialogs

Yasir Tahir*,Debsubhra Chakraborty*,Tomasz Maszczyk†,Shoko Dauwels‡,Justin Dauwels§,Nadia Thalmann* and Daniel Thalmann*

*Institute for Media Innovation, Nanyang Technological University,(yasir001,CHAK0008)@e.ntu.edu.sg

†Nicolaus Copernicus University,TKMaszczyk@ntu.edu.sg

‡College of Business, Nanyang Technological University

§Electrical and Electronics Engineering, Nanyang Technological University,JDAUWELS@ntu.edu.sg

***Abstract*—This paper proposes a real time sociometric system to analyze social behavior from audio-visual recordings of two-person face-to-face conversations in English. The novelty of the proposed system lies in this automatic inference of ten social indicators in real time. The system comprises of a Microsoft kinect device that captures RGB and depth data to compute visual cues and microphones to capture speech cues from an on-going conversation. With these non-verbal cues as features, machine learning algorithms are implemented in the system to extract multiple indicators of social behavior including empathy, confusion and politeness. The system is trained and tested on two carefully annotated corpora that consist of two person dialogs. Based on leave-one-out cross-validation test, the accuracy range of developed algorithms to infer social behaviors is 50% - 86% for audio corpus, and 62% - 92% for audio-visual corpus.**

***Index Terms*—Machine Learning, Sociometrics, Dialog, audio-visual Analysis, real-time**

I. Introduction

In an era of human centered vision of computing, social signal processing (SSP) is garnering greater interest to provide machines with social intelligence. Social intelligence is a significant component of human behavior and aids in accurate perception, interpretation and display of social signals. Non-verbal behavioral cues allow one to understand the social signals (attitudes, emotions and relations) being exchanged during conversations [1], [2]. SSP research employs sensors such as microphones and cameras to detect these cues resulting in signals (audio, visual or both) to automatically infer social interaction.

Research associated with automatic analysis of social signals includes modeling and automatic detection of personality traits, social relations and social roles from speech recordings [3] and short clips [4], [5]. In addition, automatic detection of interest in multi-party dialogs have been studied [6]. Similarly, other studies explored the inference of emphasis and interest in conversations from speech pitch [7] and the detection of agreement in meeting scenarios such as broadcast conversations [8]. Visual features were also investigated by studies that aimed at detecting dominant people and emerging leaders in group conversations [9]–[13]. It could be noted that most of these studies inferred not more than one social signal and most of the proposed algorithms have been evaluated on the annotations from one single corpus - ICSI corpus [14]. Most importantly, these algorithms are not implemented in real time.

In earlier work [15], we presented a novel approach towards comprehensive real-time analysis of speech mannerism and social behavior (interest, agreement, dominance). In this paper, we extend our work by (i) incorporating visual data along with audio data; (ii) by adding seven new sociometric measures (politeness, friendliness, frustration, empathy, respect, confusion and hostility) to further enhance our analysis [1], [2].In contrast to studies that acquired visual data using an RGB camera and determined dominance using rank and score level analysis [9]–[13], Microsoft Kinect device was employed to record RGB and depth data of the face and body of each participant from which real-time visual features were extracted. It was followed by the development of real-time algorithms (rule-based supervised learning) to automatically quantify the ten aspects of social behavior from audio and audio-visual analysis (see Fig.1). Although several corpora based on Kinect such as CAM3D and BAVCD [14], [16], [17] are available, but these are not related to human behavior. The algorithms were trained and tested on two carefully annotated corpora, namely audio corpus (AC) and audiovisual corpus (AVC) that consist of two person dialogs. We performed multi-class classification as opposed to binary classification on these corpora. Results indicate that majority of the sociometric measures can be automatically determined with reasonable accuracy (50% - 92%) in real-time, and that the combination of visual data with audio data contribute significantly to the determination of social signals. In contrast to most existing studies in social signal processing which focus on offline analysis, we are ultimately interested in developing a system that operates in real-time under real-life scenarios.

The paper is structured as follows: Section 2 focuses on data acquisition and annotation. Section 3 explains the computation of audio and video features. Section 4 presents the results of sociometric inference for audio, visual and combined modalities. Section 5 includes conclusion and future research.

II. Dataset Acquisition

In this study we generated two corpora viz AC and AVC. The AC contains 150 two-person conversations, each lasting

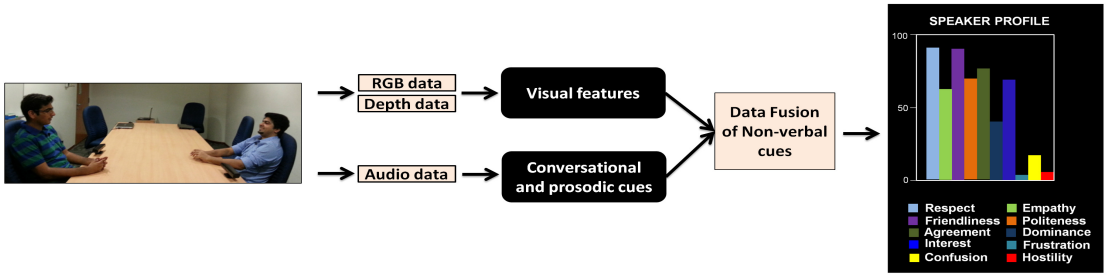


Fig. 1. System Overview.

2.5-3 minutes long. These conversations were derived from 22 students, (17 males, 5 females) who are aged between 18 and 30 years old. On the other hand, the AVC contains 100 two-person conversations, each lasting 1 minute. The total number of individuals participating in this corpus is 21, (16 males, 5 females). The participants were students of Nanyang Technological University (NTU). The topics of conversations ranged from discussion of assignments, projects of students, to social and political views. In some of the dialogs, there were scenarios such as conflicts and disagreements, periods of boredom, aggressive behavior and excessive talking. During the conversation recording procedure, the participants were seated about 1.5m apart. We saved the speech in separate channels in order to allow precise computation of overlap-related features. Kinect devices were employed to record video/depth data for each participant in AVC.

A. Annotation Protocol

Each recording was annotated by multiple judges to get consistent ratings for each sociometric. For each recording, the judges completed a brief questionnaire related to behavioral aspects of each participant. In order to annotate conversations from AC we asked the annotators to listen to the conversations, whereas for AVC they watched the video recordings and filled the questionnaire. We used a likert scale from 1 (low) to 3 (high). In the video recordings the participants could see the front profile of each speaker and could hear the audio of the left speaker on left headphone channel and vice versa. The annotators were asked to rate each clip for the sociometrics of interest, dominance, politeness, friendliness, frustration, empathy, respect, confusion, hostility and agreement. Some sample questions from the questionnaire are shown in Table 1. To assess the variability among the different annotators, we

TABLE I
QUESTIONNAIRE FOR SOCIOMETRIC ASSESSMENT.

Assessment of Social Behavior
This person seemed to be actively engaged in the conversation.
Please rate the politeness of this speaker during the conversation.
This person seemed to be the dominant of the two.

computed the standard deviation of the annotations. In Table 2 we list the minimum, maximum, mean, and median standard deviation across the 10 different sociometrics. As can be seen from Table 2, the standard deviation values are relatively low, and therefore, the annotations are reasonably consistent among the different annotators.

TABLE II
STANDARD DEVIATION VALUES FOR EACH DATASET.

Dataset	Maximum Value	Minimum Value	Mean Value	Median Value
NVAC	0.55	0.24	0.45	0.47
NVAVC	0.51	0.24	0.41	0.43

III. FEATURE EXTRACTION

In this paper, we used two different modalities of features, i.e., speech and visual. The features that we used for this analysis are shown in Table 3. In this section, we briefly review the speech and visual cues that we considered in this study.

TABLE III
LIST OF CONVERSATIONAL, PROSODIC AND VISUAL FEATURES.

Category	Features
Conversational	
Speaking Duration	Speaking % , Mutual Silence Difference in speaking %, Overlap, Response Time
Speaking Turns	Natural Turns, Turn Duration
Interruption	Interruptions, Failed Interruptions
Interjection	Interjection, Speaking Interjection
Prosodic	
Frequencies	Larynx Frequency (F0), Formant (F1, F2, F3)
MFCC	Mel-Frequency Cepstral Coefficients
Amplitude	Mean vol, max vol, min Vol, Entropy
Visual	
Postures	Upright, Hunched Forward Leaned Back, Posture Changes
Head Movement	Nodding Sum of Vertical/horizontal Head Movements
Gestures	Gesture Count, % of Gestures as compared to other speaker.
Head Pose	Straight, downward, sideward

A. Non-Verbal Speech Cues

Non-verbal speech cues play a significant role in group conversations. We briefly describe the conversational and prosodic cues as follows.

1) Conversational and Prosodic Cues: In order to compute the conversational features, we first performed speech detection by means of a hidden Markov model (HMM) that uses energy-independent features [18]. We segmented the audio signals in periods of speech and without speech, after which we computed the following conversational cues: the number of natural turns, speaking percentage, mutual silence percentage, turn duration, interjections, interruptions, failed interruptions, and response time (see Fig.2).

We considered prosodic cues such as amplitude, larynx frequency (F0), formants(F1, F2, F3) and mel-frequency cepstral coefficients (MFCCs); those cues were extracted from 30ms

segments at a fixed interval of 10ms [15]. These cues fluctuate rapidly in time. Therefore, we computed various statistics of those cues over a time period of several seconds, including minimum, maximum, mean and entropy.

B. Visual Cues

We have incorporated posture, nodding, head pose and hand gesture usage in our analysis of visual cues. The pre-processing step in evaluating the visual cues involved automatic detection of the speakers face and body as shown in Fig.2; subsequently we track them in real-time to extract visual cues.

Posture: The posture of each participant was classified into three categories: hunch forward, upright, and lean backward. We evaluated the percentage of time each participant remained in a particular posture along with total number of posture changes. Postures contribute strongly towards understanding the body language of the speaker. Since the skeleton acquired from Kinect SDK allowed us to track several points on the face and body of the speaker, we selected relevant points on the head (X_{head}, Y_{head}, Z_{head}) and neck (X_{neck}, Y_{neck}, Y_{neck}). By means of the angle θ between these points we detected the posture.

$$\theta = \tan^{-1} \frac{Y_{\text{head}} - Y_{\text{neck}}}{Z_{\text{head}} - Z_{\text{neck}}}. \quad (1)$$

Nodding: Nodding is intuitively a good indicator of agreement/disagreement between the speakers. Nodding is commonly of two types: Yes and No. We registered consecutive head movements as nodding in the algorithm. Vertical head movement corresponded to Yes and horizontal head movement to No.

Hand Gesture Detection: Instead of identifying specific hand gestures, we focused on quantifying the general hand gesture usage of the speaker. We used the number of such gestures over the course of the conversation and also the relative percentage of hand gesture usage by both speakers as features. If right hand coordinates are (X_{rh}, Y_{rh}, Z_{rh}) and left hand coordinates are (X_{lh}, Y_{lh}, Z_{lh}), we calculated the distance between these two points:

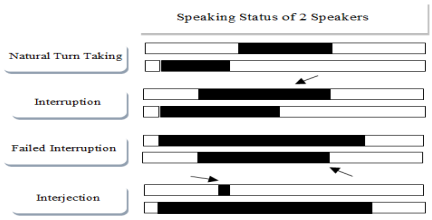
$$\text{Distance} = \sqrt{(X_{\text{lh}} - X_{\text{rh}})^2 + (Y_{\text{lh}} - Y_{\text{rh}})^2 + (Z_{\text{lh}} - Z_{\text{rh}})^2}. \quad (2)$$

If a change in this distance was above an experimentally determined threshold the algorithm recognized it as a gesture.

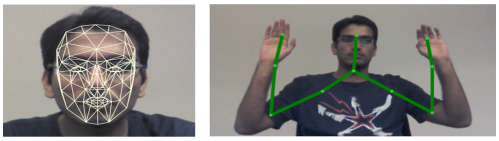
Head Pose: Face detection through Kinect SDK provides head position in right hand coordinates, and also provides yaw, pitch and roll angles to determine head pose. In our algorithm we identified the head pose using the yaw and pitch angles of the detected face. Head pose of each participant was classified as straight, sidewards and downwards. We evaluated the percentage of time each participant remained in a particular head pose.

IV. Sociometrics

This section explains how we quantify social behavior in dialogs. We explain both feature selection and the inference of sociometrics for AC and AVC.



(a) Illustration of turn-taking, interruption, failed interruption, and interjection.



(b) Detected face and skeleton.

Fig. 2. Speech features and extracted user face and skeleton from RGB-Depth data.

A. Feature Selection

We applied two feature selection algorithms, i.e., Information Gain (IG) and ReliefF [19], [20], to determine the most relevant features for inferring the sociometrics.

B. Audio Corpus

We trained multi-class classifiers in a supervised manner. The average (rounded) score provided by the judges served as labels for supervised learning and the extracted non-verbal cues were utilized as input features. In this work, we considered eight kinds of multi-class classifiers: Support Vector Machine (SVM), Support Vector Ordinal Regression, Artificial Neural Network (ANN), k-nearest neighbor, Naive Bayes, Adaptive Boosting, Bagging, Random subspace ensembles and Least squares Boosting, where the last four were ensemble classification methods [21], [22]. The classification performance was computed by leave-one-person-out cross-validation, i.e., for each participant the classifier was tested on the instances of that participant and trained on all the remaining instances. We tested both linear and RBF kernels for the SVM classifier, with parameters C for linear, C and σ for RBF. These parameters were optimized using cross-validation on training part of the data. We used the parameter values which provided the best results. Similarly for kNN the number of nearest neighbors was chosen using cross-validation. For all other classifiers we used default values from Matlab documentation. The best accuracies achieved for all the sociometrics are listed in Table 4, along with the root mean square error (RMSE) of the classifiers. Moreover, Table 4 includes the RMSE of a trivial classifier that always has the value 2 (medium) as output, which serves as a baseline for our assessment. The RMSE is each time computed between the average annotation value and the classifier output.

The numerical results in Table 4 show that agreement, dominance, interest, politeness, confusion and hostility can be detected with good accuracies from non-verbal audio features. This is not the case for friendliness, frustration, empathy and

respect, where the accuracies are quite low. The objective of this study is to determine multiple social indicators and infer human behavior based on these indicators. The results suggest that inference can be done from most of the social indicators. The low accuracies of friendliness, respect, empathy and frustration could be attributed to the absence of social hierarchy during dialogs. Since the annotators viewed both participants on the same social status, indicators such as friendliness, respect, empathy, and frustration are considered mutual. Moreover, non-verbal cues alone might not suffice to reliably infer these indicators. As expected, the RMSE values are substantially smaller than baseline for classifiers with good classification performance.

TABLE IV
THE BEST CLASSIFICATION RESULTS ACHIEVED FOR EACH SOCIOMETRIC. THE LAST TWO COLUMNS CONTAIN THE RMSE VALUES FOR ACTUAL AND BASELINE CLASSIFICATION.

Sociometrics	Audio Features	RMSE	RMSE (baseline)
Agreement	84%	0.4465	1.6503
Dominance	86%	0.3211	1.5212
Interest	85%	0.3833	1.5328
Politeness	81%	0.4920	0.8626
Friendliness	51%	0.6668	0.7280
Frustration	50%	0.7050	0.7682
Empathy	59%	0.6147	0.6460
Respect	59%	0.5939	0.7455
Confusion	81%	0.4301	0.8712
Hostility	77%	0.4935	0.9084

C. Audio-Visual Corpus

The numerical results in Table 5 suggest that the audio modality is crucial for sociometric prediction, and generally leads to better results when compared to the visual modality. In addition, combining the visual modality with audio modality significantly enhances the classification. The improvement in accuracy by including visual information is 4% for dominance, 1% for agreement, 3% for interest, 1% for politeness, 9% for friendliness, 14% for empathy and 9% for respect. For AVC data the accuracies follow a trend similar to AC. The accuracies for agreement, dominance, interest, politeness, confusion and hostility are the highest followed by the detection rates for frustration, empathy, friendliness and respect. An interesting observation from Table 5 is the considerable increase in accuracy for friendliness, empathy and respect when only visual features are used. These results indicate the significance of visual features for the inference of various social indicators.

As mentioned earlier, the best results are obtained when both audio and visual information is available, in which case the classifier trained on audio-visual data can be applied. In scenarios like call center training, where visual data is either unavailable or only occasionally available, the audio serves as the only source of information. In such cases audio-based classifiers can still provide accurate sociometric predictions, as can be seen from Table 4 and Table 5.

Our experiments show that processing time increases linearly with the duration of the conversation (see Fig.3), which makes the system scalable and feasible for real time feedback in an ongoing conversation.

TABLE V
THE CLASSIFICATION RESULTS ACHIEVED FOR EACH SOCIOMETRIC USING AUDIO, VIDEO AND AUDIO-VISUAL FEATURE SETS. THE LAST TWO COLUMNS CONTAIN THE RMSE VALUES FOR ACTUAL AND BASELINE CLASSIFICATION.

Sociometrics	Audio	Video	Audio-Visual	RMSE	RMSE (baseline)
Agreement	80%	72%	81%	0.4128	0.8504
Dominance	86%	88%	90%	0.3538	0.9715
Interest	89%	86%	92%	0.3193	0.9722
Politeness	75%	71%	76%	0.7093	0.7141
Friendliness	54%	63%	63%	0.5944	0.7689
Frustration	67%	69%	67%	0.5737	0.8915
Empathy	53%	63%	67%	0.5553	0.7168
Respect	54%	60%	62%	0.6573	0.7508
Confusion	89%	88%	89%	0.3051	0.9211
Hostility	72%	74%	72%	0.4912	0.8852

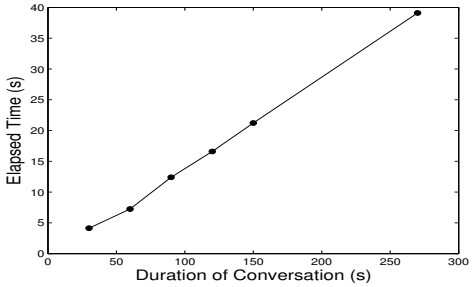


Fig. 3. This plot shows the analysis time for conversations of different durations.

V. CONCLUSION

This paper presented a novel approach towards comprehensive real-time analysis of sociometrics. Our ultimate aim is to design a real-time system that can infer social behavior during real-life conversations. We used multi-modal information (i.e. speech and video) to determine social behavior of the speaker. We collected two diverse speech corpora consisting of two-person conversations; it allowed us to train machine learning algorithms for reliable 3-level classification of the sociometrics with speech and visual cues as input features. We presented the analysis results for audio corpus (AC) and audio-visual corpus (AVC) in this paper. Results indicate that most of the social indicators can be detected with high accuracy. It also highlights the role of visual cues in enhancing the sociometric analysis. The sociometric computation is quick and reliable, enabling real-time sociofeedback. We intend to collect a much larger, diverse dataset, in order to generalize the findings.

ACKNOWLEDGEMENTS

This research project is supported in part by the Institute for Media Innovation (Seed Grant M4080824), Nanyang Technological University (NTU). Part of this study was supported by research fellowship within project Enhancing Educational Potential of Nicolaus Copernicus University in the Disciplines of Mathematical and Natural Sciences (project no. POKL.04.01.01-00-081/10).The authors thank Ms. Smitha Velayil Sunildeep for proofreading this paper.

REFERENCES

[1] Alessandro Vinciarelli, Maja Pantic, Dirk Heylen, Catherine Pelachaud, Isabella Poggi, Francesca D'Errico, and Marc Schröder, "Bridging the gap between social animal and unsocial machine: A survey of social signal processing," *Affective Computing, IEEE Transactions on*, vol. 3, no. 1, pp. 69–87, 2012.

[2] Daniel Gatica-Perez, "Automatic nonverbal analysis of social interaction in small groups: A review," *Image and Vision Computing*, vol. 27, no. 12, pp. 1775–1787, 2009.

[3] Brian Hutchinson, Bin Zhang, and Mari Ostendorf, "Unsupervised broadcast conversation speaker role labeling," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. IEEE, 2010, pp. 5322–5325.

[4] Björn Schuller, Stefan Steidl, Anton Batliner, Alessandro Vinciarelli, Klaus Scherer, Fabien Ringeval, Mohamed Chetouani, Felix Weninger, Florian Eyben, Erik Marchi, et al., "The interspeech 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism," in *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association*, 2013.

[5] Björn Schuller, Stefan Steidl, Anton Batliner, Elmar Nöth, Alessandro Vinciarelli, Felix Burkhardt, Rob van Son, Felix Weninger, Florian Eyben, Tobias Bocklet, et al., "The interspeech 2012 speaker trait challenge," in *INTERSPEECH*, 2012.

[6] Benedikt Hornler and Gerhard Rigoll, "Multi-modal activity and dominance detection in smart meeting rooms," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*. IEEE, 2009, pp. 1777–1780.

[7] Lyndon S Kennedy and Daniel PW Ellis, "Pitch-based emphasis detection for characterization of meeting recordings," in *Automatic Speech Recognition and Understanding, 2003. ASRU'03. 2003 IEEE Workshop on*. IEEE, 2003, pp. 243–248.

[8] Wen Wang, Sibel Yaman, Kristin Precoda, and Colleen Richey, "Automatic identification of speaker role and agreement/disagreement in broadcast conversation," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 5556–5559.

[9] Oya Aran and Daniel Gatica-Perez, "Fusing audio-visual nonverbal cues to detect dominant people in group conversations," in *Pattern Recognition (ICPR), 2010 20th International Conference on*. IEEE, 2010, pp. 3687–3690.

[10] Dairazalia Sanchez-Cortes, Oya Aran, Marianne Schmid Mast, and Daniel Gatica-Perez, "A nonverbal behavior approach to identify emergent leaders in small groups," *Multimedia, IEEE Transactions on*, vol. 14, no. 3, pp. 816–832, 2012.

[11] Alvaro Marcos-Ramiro, Daniel Pizarro-Perez, Marta Marron-Romera, Laurent Nguyen, and Daniel Gatica-Perez, "Body communicative cue extraction for conversational analysis," in *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*. IEEE, 2013, pp. 1–8.

[12] Roel Vertegaal and Yaping Ding, "Explaining effects of eye gaze on mediated group conversations:: amount or synchronization?," in *Proceedings of the 2002 ACM conference on Computer supported cooperative work*. ACM, 2002, pp. 41–48.

[13] Elisa Ricci and Jean-Marc Odobez, "Learning large margin likelihoods for realtime head pose tracking," in *Image Processing (ICIP), 2009 16th IEEE International Conference on*. IEEE, 2009, pp. 2593–2596.

[14] Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, et al., "The icsi meeting corpus," in *Proceedings of 2003 IEEE International Conference on Acoustics, Speech and Signal Processing, 2003 (ICASSP'03)*. IEEE, 2003, vol. 1, pp. I–364.

[15] Umer Rasheed, Yasir Tahir, Shoko Dauwels, Justin Dauwels, Daniel Thalmann, and Nadia Magnenat-Thalmann, "Real-time comprehensive sociometrics for two-person dialogs," in *Human Behavior Understanding*, pp. 196–208. Springer, 2013.

[16] Marwa Mahmoud, Tadas Baltrušaitis, Peter Robinson, and Laurel D Riek, "3d corpus of spontaneous complex mental states," in *Affective Computing and Intelligent Interaction*, pp. 205–214. Springer, 2011.

[17] Georgios Galatas, Gerasimos Potamianos, Dimitrios Kosmopoulos, Chris McMurrough, and Fillia Makedon, "Bilingual corpus for avasr using multiple sensors and depth information," in *Proc. AVSP*, 2011, pp. 103–106.

[18] Sumit Basu, "A linked-hmm model for robust voicing and speech detection," in *Proceedings of 2003 IEEE International Conference on Acoustics, Speech and Signal Processing, 2003 (ICASSP'03)*. IEEE, 2003, vol. 1, pp. I–816.

[19] Lei Yu and Huan Liu, "Feature selection for high-dimensional data: A fast correlation-based filter solution," in *ICML*, 2003, vol. 3, pp. 856–863.

[20] Igor Kononenko, Edvard Šimec, and Marko Robnik-Šikonja, "Overcoming the myopia of inductive learning algorithms with relieff," *Applied Intelligence*, vol. 7, no. 1, pp. 39–55, 1997.

[21] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar, *Foundations of machine learning*, MIT press, 2012.

[22] Wei Chu and S Sathiya Keerthi, "Support vector ordinal regression," *Neural computation*, vol. 19, no. 3, pp. 792–815, 2007.