Human and Automated Scoring of Fluency, Pronunciation and Intonation During Human–Machine Spoken Dialog Interactions

Vikram Ramanarayanan[†]*, Patrick Lange*[†]*, Keelan Evanini*[‡]*,*
Hillary Molloy[†] *and David Suendermann-Oeft*[†]

Educational Testing Service R&D,
[†]90 New Montgomery Street, Suite 1500, San Francisco, CA
[‡]660 Rosedale Rd., Princeton, NJ

vramanarayanan@ets.org

Abstract

We present a spoken dialog-based framework for the computer-assisted language learning (CALL) of conversational English. In particular, we leveraged the open-source HALEF dialog framework to develop a job interview conversational application. We then used crowdsourcing to collect multiple interactions with the system from non-native English speakers. We analyzed human-rated scores of the recorded dialog data on three different scoring dimensions critical to the delivery of conversational English – fluency, pronunciation and intonation/stress – and further examined the efficacy of automatically-extracted, hand-curated speech features in predicting each of these subscores. Machine learning experiments showed that trained scoring models generally perform at par with the human inter-rater agreement baseline in predicting human-rated scores of conversational proficiency.

Index Terms: dialog systems, computer assisted language learning, conversational assessment, intelligent tutoring systems, crowdsourcing

1. Introduction

The increasing maturation of automated conversational technologies in recent years holds much promise towards developing intelligent agents that can guide one or multiple phases of student instruction, learning, and assessment. This is very useful for applications such as language learning, since it is important to elicit learners' speech in as naturalistic a conversational setting as possible to better prepare them for the challenge of speaking a new language. Even from the assessment perspective, given that most large-scale "prompt-response" tests of non-native English proficiency (such as the TOEFL[1], BULATS[2] or Pearson Test of English Academic[3]) do not contain interactive dialogue, these types of tests are not able to elicit the full range of English speaking skills (such as turn taking abilities, politeness strategies, pragmatic competence) that are required for successful communication. Well-designed interactive dialog-based tasks have the potential to fill this gap. Spoken dialog technologies are also important since they offer opportunities for personalizing education to each learner, thereby providing a natural and practical learning interface that can adapt to a learner's individual strengths and weaknesses in real time so as to increase the efficacy of instruction [1]. In the future, such systems could potentially build an individualized profile for each learner that diagnoses gaps in knowledge and ability, adaptively composes instruction material, performs formative evaluation in many rounds of testing, scaffolds student learning through intelligent tutoring strategies, recommends when the learner is ready for a high stakes summative evaluation, and formulates long-term goals [2].

While there has been much work on the scoring of spoken monologic responses (and pronunciation scoring in particular; see [3, 4, 5, 6, 7, 8]) including the analysis of human inter-rater agreement (see [9, 10]), there has been limited work on the development and validation of automated speech scoring of spoken dialogic responses. In one study, spoken responses were collected and scored in the context of simulated dialogs [11]. In the study's design, the language learners participated in multi-turn conversations with a virtual interlocutor about university-related topics; however, the system's response was fixed and did not vary based on the learner's response, i.e., there were no branching dialog states. The conversations were then scored by expert human raters (a single score on a scale of 0-4 was provided for the learner's performance in the conversation) and automated scores were computed based on a range of speaking proficiency features aggregated across the utterances in the conversation. In another recent study, language learners' spoken responses were collected using three different task-oriented interactive dialog systems (the tasks were ordering a laptop, selecting a restaurant, and finding a bus route) [12]. Expert human raters provided proficiency scores for the learner's performance in each dialog using the CEFR scale and automated speech scoring models using features based on the audio signal and fluency characteristics were used to predict these scores. The methodology of the current study differs from these previous studies in that we obtained speaking proficiency ratings for each response in the dialog, not just for the entire conversation, and we obtained analytic ratings across multiple dimensions of speaking proficiency. The fine-grained proficiency ratings thus enable the possibility of providing more targeted automated feedback about the learner's English proficiency.

2. Data

2.1. The HALEF dialog ecosystem

We use the HALEF dialog system[4] to develop conversational applications within the crowdsourcing framework. HALEF is an open-source, modular, cloud-based dialog system that is compatible with multiple W3C and open industry standards. The HALEF architecture and components have been described

[1] http://www.ets.org/toefl
[2] http://www.bulats.org/
[3] http://pearsonpte.com/
[4] http://halef.org

Table 1: *Human scoring rubric covering three dimensions of speech delivery (fluency, pronunciation, and intonation/stress)*

	4 Very Good	3 Generally good	2 Somewhat Limited	1 Very Limited	0
1. Fluency	*Very good tempo and minimal hesitation.* The response includes pauses at appropriate locations to formulate ideas.	*Good tempo and minimal hesitation*. The response includes some pauses to formulate ideas which minimally impacts the flow of speech.	*Noticeable pauses and hesitations*. The tempo is choppy, and/or filler words are frequent in the response.	*Frequent long pauses and/or use of filler words*. It is challenging to follow the flow of ideas due to frequent long pauses and/or filler words.	No response or no English in the response
2. Pronunciation	*Highly intelligible*. Though the response may include L1 influence, word-level pronunciation do not impact intelligibility.	*Generally intelligible*. Though the response may show noticeable L1 accent, word-level pronunciation do not significantly impact intelligibility.	*Generally unintelligible*. The response shows noticeable L1 accent. Errors in word-level pronunciation occasionally hinder intelligibility.	*Unintelligible*. The response shows noticeable L1 accent. Errors in word-level pronunciation substantially impact intelligibility.	OR The response is completely unrelated to the test
3. Intonation/Stress	*Appropriate sentence-level intonation and stress used to convey meaning*. Intonation and stress do not hinder intelligibility.	*Generally appropriate sentence-level intonation and stress used to convey meaning*. Non-target intonation and stress may mildly impact intelligibility.	*Generally inappropriate sentence-level intonation and stress used to convey meaning*. Non-target intonation and stress impact intelligibility.	*Inappropriate sentence-level intonation and stress used to convey meaning*. Inappropriate intonation and stress significantly reduce intelligibility.	OR Non-scorable (e.g., audio file is largely unintelligible)

in detail in prior publications [13, 14, 15].

2.2. Crowdsourcing data collection

We leveraged the aforementioned HALEF dialog system to develop conversational applications within an Amazon Mechanical Turk crowdsourcing framework. In this iterative data collection framework, the data logged to the database during initial iterations is transcribed, annotated, rated, and finally used to update and refine the conversational task design and models for speech recognition and spoken language understanding. Since the targeted domain of the tasks in this study is conversational practice for English language learners, we restricted the crowdsourcing user pool to non-native speakers of English. In all, we collected 768 conversational interactions between callers and the dialog system for the interview task described below, 123 of which were used as the cross-validation train/test set for automated scoring experiments (1893 utterances, $17,456$ tokens; average turn length = $6.3s$; average call duration = $364s$) the remaining 645 were used as a development set (8672 utterances, 83755 tokens; average turn length = $6.5s$; average call duration = $304s$) for tuning the automatic speech recognizer language models used in the speech scoring experiments[5].

2.3. The Conversational Interview Task

This study examines a conversational task developed for English language learners that was designed to provide speaking practice for non-native speakers of English in the context of a simulated job interview. The conversation is set up as a system-initiated dialog in which a representative at a job placement agency interviews the language learner about the type of job they are looking for and their qualifications. The ultimate aim of the task is to provide interactive feedback to language learners about whether they have demonstrated the linguistic skills necessary to provide appropriate, intelligible responses to the questions and to complete the communicative task successfully.

2.4. Scoring

In order to understand how well participants performed in our conversational interview task, we had two human scorers listen to participant responses corresponding to each dialog turn in the train/test set and score them on the following three dimensions according to the rubric shown in Table 1: fluency, pronunciation and intonation/stress. The scoring procedure was as follows: both raters first provided independent ratings for responses from a small number of conversations and then reviewed their ratings together to reach a consensus about any discrepant ratings. They then proceeded to apply that scoring process to the remaining conversations in the cross-validation set. A subset of 33 conversations (365 utterances) were double-scored; the remaining 90 conversations (1528 utterances) in the cross-validation set received a score from one of the two raters.

[5]We partitioned the dataset in this manner in order to have as much data as possible available for training robust language models

3. Scoring Analyses and Experiments

This section first presents an analysis of how well non-native English speakers performed on our deployed conversational task. We then examine how well features that are currently used in automated speech scoring research (covering diverse measurements among lexical usage, fluency, pronunciation, prosody, etc.) fare in automatically predicting human-rated scores of speaker performance.

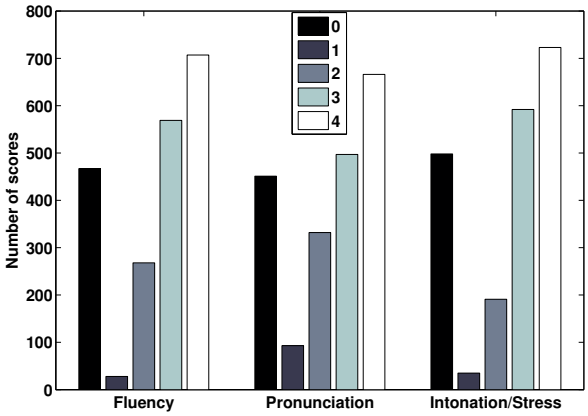

Figure 1: *Distributions of human ratings*

3.1. Human Scoring Analysis

Figure 1 plots the distributions of various sub-scores assigned by human raters (average ratings are presented for the responses

that were double-scored). The scores are mainly distributed across the $2-4$ range, with relatively few scores of 1. Also note that a large number of utterances (>400 per sub-score dimension) were marked unscorable (score of 0) owing primarily to poor audio quality or unintelligible speech responses.

Table 2 lists the inter-rater correlations ($\rho_{R_1R_2}$) and quadratic weighted kappa (κ) values computed on the scores assigned by both our two human scorers (365 utterances). Notice that the $\rho_{R_1R_2}$ value for the three scoring dimensions lie between 0.54 and 0.61. The overall inter-rater agreement, as measured by κ, computed across all scoring sub-dimensions, was 0.49 (although this was not significant at the $\alpha = 95\%$ level). This indicates a moderate level of agreement between our two raters, but more importantly, this underscores two possibilities. First, scoring non-native dialogic responses is non-trivial and potentially more challenging as compared to monologic counterparts due to a variety of possible reasons, including, but not limited to, the increased spontaneity of responses, greater occurrence of disfluencies, dependence of the learner's responses on the conversation history, and the short duration of each individual response. Second, the scoring rubric we used was not optimized to this particular dialogic task. Having said that, this type of study is an important first step towards developing a refined rubric for scoring dialog constructed responses since there is no defined state-of-the-art rubric for scoring such data. Going forward, the inter-rater agreement could be potentially improved with additional rater training and calibration and having more (than two) raters score the data.

3.2. Automated Scoring

With an understanding of the score distributions, we now move on to experiments in automatically assigning such scores using machine learning techniques. Such methods could play an important role in dialog management routines that direct and adapt conversation flow, and eventually, toward automating the process of scoring. Therefore, here we train regression models to predict each of the average human sub-scores (listed in Table 1) assigned to all spoken constructed responses in our train/test set in a cross-validation setup[6].

For automatic scoring of these spoken constructed responses at each dialog turn, we extracted features that are currently used in automated speech scoring research, covering diverse measurements among lexical usage, fluency, pronunciation, prosody, and so on. In particular, following the feature extraction method described in [16], we used SpeechRater^{SM}, a speech scoring system that processes the speech signal and its corresponding ASR output to generate a series of features assessing multiple dimensions of speaking proficiency, e.g., speaking rate, prosodic variations, pausing profile, pronunciation, and vocabulary diversity. In total, 144 SpeechRater features were included in this experiment. In the interest of brevity, we do not discuss them in detail here; for more on these features, refer to [16, 5, 17, 10]. The majority of SpeechRater features leverage information derived by running the input audio files through an automatic speech recognizer (ASR). It therefore follows that in order to optimize the feature extraction process, we need to optimize the models that the ASR leverages, in particular, its acoustic and language models. We trained our acoustic model on 800 hours of non-native speech data collected from TOEFL iBT test administrations. We used standard 13-dimensional MFCCs with deltas and delta-deltas and 10ms shift. The final acoustic model is a p-norm DNN [18] with 4 hidden layers, a dimensionality of the input/output layer of 2000/250 and was trained in 8 epochs. The systems phonetic alphabet is comprised of 42 basic tokens combining 39 true phonemes, and tokens for "silence", "spoken noise" and "noise". Additionally, the final phonological tokens have word position-specific modifiers for internal, singleton, word-beginning and word-ending positions. We used a held out training set to train a 3-gram language model using modified Kneser-Ney smoothing. The model has a perplexity of 46 on the test data.

[6] As described in Section 2.4, not all utterances received double human scores; in these cases, only a single human score was used for training the model, not the average of two scores.

Table 2: *Scoring prediction results for different machine learning regression models relative to the human agreement in terms of quadratic-weighted Cohen's kappa, κ and the inter-rater correlation, $\rho_{R_1R_2}$ (Classifer acronyms are as follows: EN = Elastic Net; RF = Random Forest Regressor; GB = Gradient Boosting Regressor).*

Scoring Dimension	κ	$\rho_{R_1R_2}$	Model Correlation		
			EN	RF	GB
Fluency	0.48	0.54	0.51	**0.58**	0.55
Pronunciation	0.54	**0.61**	0.51	0.56	0.54
Intonation/Stress	0.45	**0.57**	0.47	0.54	0.51

3.3. Machine Learning Experiments

We used SKLL,[7] an open-source Python package that wraps around the *scikit-learn* package [19] to perform machine learning experiments. We experimented with a variety of learners to predict the various delivery scores, including regularized linear regressors (Lasso [20] and Elastic Net [21]), tree-based regressors (e.g., Random Forests [22]), and boosting-based learners (e.g., Gradient Boosting [23, 24]), and report the results of the best performing regression model from each learner type. We used quadratic weighted kappa (which takes into account the ordered nature of the categorical labels) as an objective function for optimizing learner performance. We further tuned and optimized the free parameters of each learner using a grid-search method. We ran stratified 10-fold crossvalidation experiments, where folds were generated to preserve the percentage of samples in each class. The experiments looked at only audio files at the level of each dialog turn. Also, note that we only examined data from individual utterances that were assigned ratings between 0 and 4.

3.4. Observations and Results

As shown in Table 2, we generally observed that Random Forest learners significantly outperform regularized linear regressors such as the Elastic Net[8] across the board. In addition, the performance of the automated scoring models were generally comparable to the inter-rater correlations,[9] $\rho_{R_1R_2}$, indicating that the SpeechRater features are able to capture the speaking proficiency characteristics in the rubrics that were applied by

[7] https://github.com/EducationalTestingService/skll

[8] The Elastic Net generally outperformed other learners in this category such as Support Vector Regression or Lasso (which is a special case of the Elastic Net with a zero L2 regularization or ridge regression penalty term) and hence we only show the performance of the Elastic Net here.

[9] Note that the sample sizes for the model correlation and inter-rater correlation results differ, since not all responses were rated by two raters, as described in Section 2.4.

Table 3: *Top 5 features which are most correlated with scores (excluding* 0 *scores).*

	Fluency		Pronunciation		Intonation/Stress	
	Feature selected	ρ	Feature selected	ρ	Feature selected	ρ
1	Duration-normalized number of types	0.34	Duration-normalized number of types	0.20	Duration-normalized number of types	0.17
2	Number of words per second	0.30	Content vector analysis score	0.19	Number of words per second	0.15
3	Duration-normalized number of types (excluding pauses and disfluencies from duration)	0.24	Unweighted average confidence score	0.19	Duration-normalized number of types (excluding pauses and disfluencies from duration)	0.11
4	Acoustic model score	0.22	Confidence score per second	0.18	Content vector analysis score	0.09
5	Average rank of word types in the response	0.21	Type to token ratio	0.17	Proportion of types in response vs. reference list of academic vocabulary	0.08

the human raters.

In order to understand why SpeechRater features performed well, we examined the five features that were most correlated with the delivery scores along each dimension (see Table 3). We observe that delivery-oriented features such as the duration-normalized number of types and word rate, as well as the acoustic model score, are highly correlated with fluency scores. However, while the confidence score-based features are potentially useful measures of pronunciation, it is not clear that the other features *directly* inform pronunciation or intonation subscores. In these cases, it is likely that different sub-scores are likely to be correlated, allowing one to achieve a reasonable predictive power by measuring aspects of proficiency which are not actually relevant to that subscore.

4. Discussion

While the results presented are very encouraging, it is important that we also look at some open questions and limitations. This study has looked at one conversational task – a job interview – in depth, and has demonstrated the efficacy of the HALEF dialog-based framework in eliciting speech from non-native English speakers while maintaining favorable user experience. However, one should note that this is but one version of the task that caters to speakers of a moderate conversational proficiency level, and may not be as useful to learners at other levels, novice speakers, for example. For this reason, we aim to broaden the scope of the interview task going forward, for instance, by creating multiple versions of the task to cater to populations at different proficiency levels. Also, we have analyzed only one task in detail here – and while that is already an important step considering the lack of related work in this particular area – in order to get a better idea of how people respond to different scenarios we will need to deploy and analyze more conversational tasks, and collect data from a larger sample of language learners.

Another important set of observations have to do with the rubric used for scoring. As we pointed out earlier, while the rubric used was designed for long constructed spoken responses, it does not explicitly consider the dialogic nature of many conversational responses. In other words, the length, complexity, and appropriateness of each response may depend on the conversation history to varying extents. This is a factor we will consider in designing rubrics for future scoring research.

For automatic scoring and rating prediction experiments, we looked at three primary classes of learners – regularized linear, tree-based and boosting-based. While these were chosen primarily for interpretability (which becomes more crucial for high-stakes assessment) and for the sake of comparison with existing state of the art in automated scoring, this particular study has not examined how well cutting-edge deep neural networks (DNNs) perform on the same task. While these networks suffer from the drawback of being relatively uninterpretable, they have significantly outperformed other machine learning methods on tasks in the speech, vision and language processing community. For this reason, we will look at how well DNNs perform on automatic scoring and rating tasks as part of future work.

5. Summary and outlook

This paper has presented a methodology and framework for computer assisted language learning of conversational English based on spoken dialog. The framework leverages the HALEF open-source modular standards-compliant dialog system deployed in a crowdsourcing data collection paradigm to obtain conversational data from potential English language learners, i.e., non-native speakers of English. We had human raters score the utterances on three dimensions representing speech delivery that are important in conversational proficiency assessment. We then extracted curated speech features based on the SpeechRater engine, and showed that these features, when fed into trained machine learning models, are able to automatically predict the human scores at a level that is comparable to human agreement.

Future research and development will look at standardizing and extending the analyses presented in this paper to a wider variety of conversational tasks and non-native speakers of varying proficiency levels. Technology-wise, we would like to make the conversational tasks truly multimodal, with support for video and text, and perhaps simulated environments and virtual avatars, to allow for a more immersive interactive experience. Yet another area of future work is the iterative refinement of the scoring rubrics to better suit dialog and conversational data, such that we adequately capture the various sub-dimensions that contribute to a language learner's conversational speaking proficiency.

6. Acknowledgements

We are grateful to Saerhim Oh, Larry Davis, Veronika Timpe-Laughlin, Tanner Jackson, Pablo Garcia Gomez, John Norris, and Spiros Papageorgiou for creating the scoring rubric and contributing to the conversational task design. We would like to also thank Lydia Rieck, Elizabeth Bredlau, Katie Vlasov, Juliet Marlier, Phallis Vaughter, and Nehal Sadek for their help in designing the conversational tasks, and Robert Mundkowsky, Ben Leong, Chong Min Lee and Dmytro Galochkin for their engineering support. Finally, we thank Juan Manuel Bravo for his help in patiently scoring all the dialog turn audio data.

7. References

[1] A. Kerry, R. Ellis, and S. Bull, "Conversational agents in e-learning," in *Applications and Innovations in Intelligent Systems XVI*. Springer, 2009, pp. 169–182.

[2] A. Graesser and B. McDaniel, "Conversation agents can provide formative assessment, constructive learning, and adaptive instruction," *The future of assessment: Shaping teaching and learning. Mahwah, NJ: Erlbaum*, 2007.

[3] L. Neumeyer, H. Franco, V. Digalakis, and M. Weintraub, "Automatic scoring of pronunciation quality," *Speech communication*, vol. 30, no. 2, pp. 83–93, 2000.

[4] S. M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech communication*, vol. 30, no. 2, pp. 95–108, 2000.

[5] K. Zechner, D. Higgins, X. Xi, and D. M. Williamson, "Automatic scoring of non-native spontaneous speech in tests of spoken english," *Speech Communication*, vol. 51, no. 10, pp. 883–895, 2009.

[6] X. Xi, D. Higgins, K. Zechner, and D. Williamson, "A comparison of two scoring methods for an automated speech scoring system," *Language Testing*, vol. 29, no. 3, pp. 371–394, 2012.

[7] S. Bhat and S.-Y. Yoon, "Automatic assessment of syntactic complexity for spontaneous speech scoring," *Speech Communication*, vol. 67, pp. 42–57, 2015.

[8] A. Loukina, K. Zechner, L. Chen, and M. Heilman, "Feature selection for automated speech scoring," in *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, 2015, pp. 12–19.

[9] X. Xi and P. Mollaun, "Investigating the utility of analytic scoring for the toefl academic speaking test (tast)," *ETS Research Report Series*, vol. 2006, no. 1, 2006.

[10] L. Chen, K. Zechner, S.-Y. Yoon, K. Evanini, X. Wang, A. Loukina, J. Tao, L. Davis, C. M. Lee, M. Ma, R. Munkowsky, C. Lu, B. Leong, and B. Gyawali, "SpeechRater 5.0," *ETS Research Report Series*, in press.

[11] K. Evanini, S. Singh, A. Loukina, X. Wang, and C. M. Lee, "Content-based automated assessment of non-native spoken language proficiency in a simulated conversation," in *Proceedings of the Machine Learning for SLU & Interaction NIPS 2015 Workshop*, 2015.

[12] D. Litman, S. Young, M. Gales, K. Knill, K. Ottewell, R. van Dalen, and D. Vandyke, "Towards using conversations with spoken dialogue systems in the automated assessment of non-native speakers of english," in *17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 2016, p. 270.

[13] V. Ramanarayanan, D. Suendermann-Oeft, P. Lange, R. Mundkowsky, A. V. Ivanov, Z. Yu, Y. Qian, and K. Evanini, "Assembling the Jigsaw: How Multiple Open Standards Are Synergistically Combined in the HALEF Multimodal Dialog System," in *Multimodal Interaction with W3C Standards*. Springer, 2017, pp. 295–310.

[14] Z. Yu, V. Ramanarayanan, R. Mundkowsky, P. Lange, A. Ivanov, A. W. Black, and D. Suendermann-Oeft, "Multimodal HALEF: An Open-Source Modular Web-Based Multimodal Dialog Framework," in *Proc. of the IWSDS Workshop 2016, Saariselka, Finland*, 2016.

[15] D. Suendermann-Oeft, V. Ramanarayanan, M. Teckenbrock, F. Neutatz, and D. Schmidt, "Halef: An open-source standard-compliant telephony-based modular spoken dialog system: A review and an outlook," in *Natural Language Dialog Systems and Intelligent Assistants*. Springer, 2015, pp. 53–61.

[16] L. Chen, K. Zechner, and X. Xi, "Improved pronunciation features for construct-driven assessment of non-native spontaneous speech," in *Proceedings of NAACL-HLT*, 2009.

[17] D. Higgins, X. Xi, K. Zechner, and D. M. Williamson, "A three-stage approach to the automated scoring of spontaneous spoken responses," *Computer Speech and Language*, vol. 25, no. 2, pp. 282–306, 2011.

[18] X. Zhang, J. Trmal, D. Povey, and S. Khudanpur, "Improving deep neural network acoustic models using generalized maxout networks," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 215–219.

[19] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[20] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[21] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

[22] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[23] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.

[24] ——, "Stochastic gradient boosting," *Computational Statistics & Data Analysis*, vol. 38, no. 4, pp. 367–378, 2002.