Parkinson's Disease Detection- An Interpretable Approach to Temporal Audio Classification

Raj Shah*, Bhavi Dave†, Nirali Parekh‡ and Kriti Srivastava§
Department of Computer Engineering
Dwarkadas J. Sanghvi College of Engineering
Mumbai, India
*rajshah2320@gmail.com, †bhavidave5@gmail.com,
‡nirali25parekh@gmail.com, §kriti.srivastava@djsce.ac.in

2022 IEEE 3rd Global Conference for Advancement in Technology (GCAT) | 978-1-6654-6855-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/GCAT55367.2022.9971881

***Abstract*—Model Interpretability is critical for analyzing the potential risks inherent to the utilization of black-box neural networks in the medical domain. The current methodologies for Parkinson's disease (PD) detection require expertise from medical professionals alongside an expensive and time-consuming procedure. AI-driven Deep Learning (DL) models have generated state-of-the-art performance and have the potential to improve the process of PD detection by leveraging the early stage symptom of vocal deterioration to enable early identification of the disease. However, the potential of DL for PD selection has remained untapped thus far because of the lack of model interpretability and the possibility of even an accurate model failing to capture the correct causal relationship between the input signal and the prediction. This work implements a neural network that contains vowel phonations. The model achieved a classification accuracy of 90.32% with a precision, recall and F1 Score of 0.91, 0.90 and 0.905. A novel model interpretation algorithm is proposed that divides the audio file into logical chunks and provides a category label identifying the productivity of each chunk towards correct inference. This is based on its interactions with every other audio region in the file, and the effective change of model performance and confidence with its inclusion. An audio instance within a clip is significant for desired output only in conjunction with other audio regions that together generate a meaningful pattern indicating the existence of PD. The algorithm also ranks all audio chunks in the order of their importance to the correct prediction by analyzing the degree in which their impact varies when combined with other audio regions. This paper demonstrates that DL can be leveraged to create a reliable, accurate, and efficient method for PD Detection and the model can be extended to provide interpretability to the inferences in a way that is scalable, and free from bias.**

***Index Terms*—Parkinsons Disease, Explainable AI, Deep Learning, Spectrograms, Convolutional Neural Network**

I. Introduction

ML techniques, despite their success, have several limitations and drawbacks. [1] The poor explainability of ML complicated models, for example, has limited their widespread use in sectors where the consequences of a choice is essential, such in medical applications.

Explainable AI is a technique wherein the solution's findings are human-understandable. It is different from the traditional "black box" concept in machine learning, where even its developers are not able to explain why their Artificial Intelligence Model made a certain decision. [2]. Physicians, on the other hand, should not rely solely on the model's predictions, but also on the findings obtained. Physicians must comprehend the reasoning and methodology with which these decisions were made in order to trust them, and compare these reasoning to their previous domain expertise.

The number of individuals affected by Parkinson's disease (PD) is increasing as the world's population ages. If explainable methods could be implemented to a speech recording dataset to accurately detect Parkinson's disease, it would be a useful screening tool before seeing a clinician. [3] The number of persons affected by 1 7. References Parkinson's disease (PD) is increasing. Efficient and early PD detection is hindered beacuse of a lack of facilities and knowledge.The signs of all PD patients are varied. [4], [5]. Speech and voice pattern analysis techniques might be useful in detecting Parkinson's disease early on. There is a lack of any laboratory biomarkers for Parkinson's disease, and brain imaging studies do not provide a conclusive diagnosis. These criteria have a 90% accuracy rate for detecting Parkinson's disease, however it takes an average of 2.9 years to get a diagnosis. The early identification of Parkinson's disease, as well as the prediction of therapy, would have a significant impact on both patient quality of life and the health service [6].

II. Motivation

Medicine has long reached an overwhelming consensus on the importance of detecting Parkinson's disease in a timely manner. With rapid developments in the domain of medical technologies , Deep Learning approaches are able to achieve state of the art performance in solving several problems in the medicinal arena. [7]

Our architecture is modeled on a typical black-box neural network and extends the architecture to provide interpretability to model the aforementioned model inferences. This would overcome the shortcomings of the traditional method of diagnosis where a physician is required to perform a tedious analysis of a person's motor skills in various situations. Vocal deterioration is among the first symptoms of the disorder and accurate and reliable models for the same can help with early detection and therefore promote better treatment. Our model can serve as an effective non-invasive screening tool. As deep learning continues to be adopted, the prominence of assistance and automated decisions made by neural networks in high

978-1-6654-6855-8/22/$31.00 ©2022 IEEE

stake situations is an undeniable fact that stakeholders and academicians have to grapple with [8]. An XAI model that detects PD leverages the performance of a neural network while providing reasoning and accountability to inferences, reducing bias. Consequently, our algorithm can help the adoption of neural networks for Parkinson's disease identification by proving interpretability and therefore reasoning for model inferences. The potential for human scrutiny, interpretation and verification can make the model robust and reliable, increasing confidence.

III. Literature Review

Research in the realm has been focused on leveraging the use of speech recognition to resolve the issue of Parkinson's Disease prediction. DeepVoice (H. Zang et al) [9] is among the most prominent works and proposes a voiceprint-based framework by integrating Deep Learning in the mobile health domain for PD identification. They use a CNN for PD final identification. (Ondřej Klempíř et al.) [10] uses Machine Learning on speech utterances for detection. Since there are natural variations in biomarkers like articulation and speech, several ML methodologies like bagged trees, SVMs (Quadratic), KNN and AdaBoost were used.

Another approach is the use of transfer learning to generate a robust and automated PD detector that also uses audio signals for identification [11]. Klempir et al. trained Convolutional neural networks (CNN) like SqueezeNet1.1, ResNet101, and DenseNet161 for automated PD identification using the natural biomarkers present in voice signals. O Karaman et Al [12] proposed novel feature engineering methods alongside established ML techniques.

Another novel approach leveraged variational mode decomposition on speech audio signals for PD diagnosis. M Yang et Al. [13] used the aforementioned technique to gain relevant signals from the entirety of audio data. In [14], K Biswajit et al forth four ML methodologies and eighteen feature extraction techniques [15]. The combination of speech signal processing with novel and traditional methodologies is extremely effective in extracting "dysphonia features" that could be both linear and non-linear in nature. These automatic and accurate identifications are sophisticated and accurate. [16], [17].

TABLE I: Literature Review Analysis

Ref	Methodology	Performance
[11]	CNN + Joint Time-Frequency Analysis algorithm	Accuracy : 90.45 ± 1.71%
[18]	AdaBoost, Bagged trees, Quadratic SVM and k-NN	Average overall accuracy : 82.3% , Averaged AUC = 0.88
[13]	DenseNet-161 (CNN) + transfer learning	Accuracy: 89.75% , Sensitivity: 91.50%, Precision: 88.40%
[14]	SVM + 10-fold cross-validation (CV)	Accuracy = >58%
[15]	Variational Mode Decomposition (VMD)	Accuracy: 96.29%
[16]	18 feature extraction techniques + 4 machine learning methods	Accuracy: 94.55%, AUC 0.87, EER 19.01%.
[17]	Novel speech signal processing + classical regression and classification algorithms	Accuracy: 88.74%, Recall: 97.03%, Mean absolute error: 3.7699
[4]	CNN + DNN with time-frequency image features	Accuracy: 90%
[19]	Stacked autoencoder and a softmax classifier	Accuracy: 86.095%
[12]	AdaBoost, Bagged trees, Quadratic SVM & k-NN	Accuracy: 82.3% , Average AUC: 0.88

A. Research Gaps

1) There is an imminent lack of explainable algorithms that are designed specifically for audio since the current literature of XAI architectures does not focus on time series signals like audio.
2) Even relevant XAI and model interpretability methodologies do not generalize across varied model architectures that can be used to analyze similar signals.

IV. Data Description

In this work, the dataset named "Synthetic vowels of speakers with Parkinsonism" contributed in 2019 is used for experimentation [9]. The dataset contains synthesized replicas of sustained vowels ‘A’ and ‘I’ performed by healthy controls and patients with Parkinson’s disease.

Figures 1 (a) and (b)reveal some differences between the speech of healthy subjects and Parkinson’s patients. Manual inspection of the audio files exhibited that the subjects with Parkinson’s delivered slurred words, mumbling, or even stuttering. Past research also shows that 89% of people with Parkinson’s disease (PD) experience speech and voice disorders.

Fig. 1: Sample Audio Waveform of a healthy person

Fig. 2: Sample Audio Waveform of Parkinson’s’ patient

A spectrogram is a visual representation of the spectrum of frequencies of a signal as it varies with time. Figures 2 (a) depicts a spectrogram of an audio file.

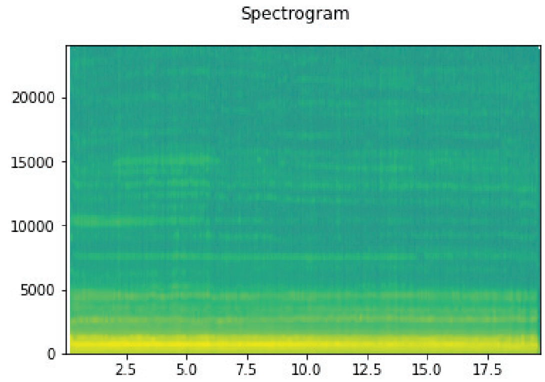

Fig. 3: Sample Spectrogram

V. Architectural Design

The proposed algorithm is a model agnostic surrogate model that leverages the variance in changes in the prediction to interpret black-box neural networks for audio classification. The trivial and simplified version of the model intuition is that if the model prediction does not change much by tweaking or removing the value of an audio chunk, the particular audio chunk is likely not an important predictor because its effect on the desired model output is negligible. [20]

This hypothesis is further developed to include cases where a particular audio instance within a clip is significant for desired output only in conjunction with other audio regions that together generate a meaningful pattern indicating the existence of PD. That is, an audio chunk might not alone be indicative of the ailment, but when seen in the context of other audio regions from the same clip, indicate the same. It is for this reason that the model output cannot be attributed to a single audio chunk without analyzing it with all the other audio chunks.

Therefore, a model cannot attribute importance to an audio region solely based on its presence or absence - we must take into account all the possible interactions with other audio regions that lend credence to a particular datapoint. The proposed algorithm aims to generate a sophisticated model of these interdependencies in a way where the true contribution of each audio region can be estimated, along with the confidence of the estimation.

A. Methodology

A black-box deep neural network was trained on the dataset. A convolutional neural network (CNN) was selected because it gains relevant, abstract, and location invariant features while the max-pooling layer helps reduce complexity while allowing prominent features to pass through to the upper layers of the network [21].

The audio files are read as a tensor and resampled to a fixed sample rate of 22050 kHz. The time-series signal is then transformed into the image domain by converting it into a spectrogram that represents the frequency content in the audio file as image colors. The work employs a log-scaled Mel-spectrogram [22] that converts frequencies to the mel scale that takes into account human sensitivity to different ranges of audio frequency. This neural network is therefore trained to perform binary audio classification - given an audio file, it predicts whether the speaker belongs to the Parkinson’s disease class or the healthy control class. The model was trained on the same partition of the dataset to avoid overlap, ensuring that the test data points were unseen by all three models. 15% of the dataset was reserved for testing, and the same other 80% was used to train the model. Stochastic gradient descent was used as an optimizer with a momentum of 0.9. The criterion used to evaluate loss for backpropagation was Cross-Entropy Loss [23] .

After the neural network is trained, the proposed algorithm for model interpretability is applied. Every audio file is sampled and a tensor from every second of audio is collected, standardizing the representation of one second of audio as an “audio chunk”. The algorithm can provide interpretability to model predictions for audio of any given length, because of the aforementioned standardization.If audio is not large enough to fill a chunk, zeros are padded at the trailing end to maintain the generalization. The pipeline is shown in the Figure 3.

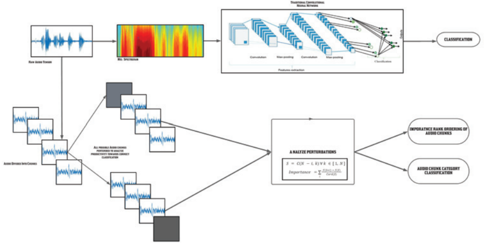
Raw Audio Tensor
Mel Spectrogram
Traditional Convolutional Neural Network
Convolution
Max-pooling
Convolution
Max-pooling
Classification
Features extraction
Outputs
CLASSIFICATION
Audio Divided Into Chunks
All possible Audio chunks perturbed to analyze productivity towards correct classification
A NALTZE PERTURBATIONS
$S = C(N - t, k) \forall k \in [1, N]$
IMPORATNCE RANK ORDERING OF AUDIO CHUNKS
AUDIO CHUNK CATEGORY CLASSIFICATION

Fig. 4: Audio Chunk Category Classification

Since every region of audio can interact with every other audio region of varying size, the interpretable algorithm considers each of these interactions and their effect on the black-box neural network's output. While predicting masked the class of an audio file with a masked chunk, there are 6 cases possible as is highlighted in Table II.

TABLE II: Category Classification for Audio Chunks

Case	Initial Prediction	Prediction after Masking	Change in Confidence	Category
1.	Correct	Correct	Higher	Unproductive
2.	Correct	Correct	Lower	Marginally productive
3.	Incorrect	Incorrect	Higher	Productive
4.	Incorrect	Incorrect	Lower	Marginally unproductive
5.	Correct	Incorrect	NA	Crucially Productive
6.	Incorrect	Correct	NA	Extremely Unproductive

We can therefore immediately label every audio second as productive or unproductive with a granularity of 6 classes.

We define the productivity of an audio chunk as the percentage difference in the confidence of the correct class when the section of audio is excluded.

$$\text{Productivity of a Chunk} \;=\; \frac{\text{NewC - OgC}}{\text{OgC}} \tag{1}$$

$$where,\ NewC = New\ Confidence$$

$$of\ correct\ class$$

$$OgC = Original\ Confidence$$

$$of\ correct\ class$$

Since every second of audio data does not enhance or degrade performance in isolation, the delta in model performance is attributed to its interaction with any other second in the audio in groups of all possible sizes. That is, the importance of an audio chunk is calculated based on its productivity value calculated in association with unique combinations of sizes ranging from one to the audio length. [24] The algorithm, therefore, considers all possible interactions and combines the productivity results in all cases. Before aggregation, these values are normalized by the number of audio data points in the group (i.e. the cardinality of S) since the deviance in performance can be attributed to all the chunks in the group.

$$\text{Importance} = \frac{1}{\text{Card}(S)} \sum_{\forall S} \text{Productivity}(S) \tag{2}$$

Consequently all the audio chunks that compose the audio file can be ranked according to the black-box neural network's interpretation of their productivity, and consequently, their importance. The deviation in these values for every second shows the consistency with which the interpretable algorithm produces the results and can therefore be attributed to the XAI model's confidence in the productivity value. The most productive seconds are the most important ones.

B. Algorithm

If N is the length of the Audio file, f is a probability function demonstrating the output of a black-box neural network, S denotes sunsets and C denotes combinations then the importance of second i can be calculated in by iteratively forming subsets and aggregating a weighted importance value normalized by the cardinality of the subset. [25]

$$S = C(N - i, k) \forall k \in [1, N] \tag{3}$$

$$\text{Importance} \;=\; \sum_{i} \frac{f(S+i) - f(S)}{\text{Card}(S)} \tag{4}$$

The algorithm, therefore, provides interpretability to a black box neural network output.

VI. Results

This section outlines the evaluation measures and the results obtained by the black-box neural network as well as the interpretability provided by the novel algorithm. The results are summarized in the Table III. All metrics are calculated on unseen test data.

TABLE III: Result Analysis

Metric	Model Performance
Accuracy	90.32%
Precision	0.91
Recall	0.90
F1 score	0.905

Furthermore an example of the per chunk analysis performed by the model can be seen in the Figure 4. As described in the methodology section, the model-interpretability algorithm considers all possible interactions between subset combinations of audio chunks and generates a category label for each audio chunk using majority voting depending on the weighted values of change in model performance.

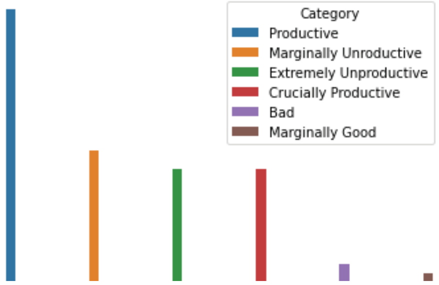

Fig. 5: Audio Chunk Category Classification

As described in the methodology section, for every 'n' second audio, the model also ranks all audio chunks in the order of importance and associates a confidence value associated with the rank. This helps determine the relative importance of audio chunks and the degree by which their impact varies when combined with other audio regions. An example output for a 10 second audio can be seen below in Table IV -

TABLE IV: Rank Ordering of Audio Chunks

Audio Chunk Number	Importance Rank	Confidence
1	1	0.82
4	2	0.71
7	3	0.54
2	4	0.26
9	5	0.84
0	6	0.91
6	7	0.76
8	8	0.79
5	9	0.46
3	10	0.53

VII. Conclusion and Future Scope

Our project has implemented a novel, interpretable algorithm. Given voice recordings of characteristic vocal features, we predict if a voice sample belongs to the set of patients having Parkinson's or if it is amongst the group of healthy control. This non-invasive tool for diagnosis can help promote early detection and treatment, overcoming the tedious traditional method. Adding explainability can help the implementation of the algorithm at scale, especially in a sensitive domain like medicine. Since we can follow the model's chain of reasoning, the likelihood of bias is low and is susceptible to human verification. Our novel XAI approach will be among the first Explainable architectures for audio and does not place restrictions on the input audio data, coping with variable lengths. Our method of modeling Sequence data and revising traditional sequence models are highly generalizable and can be implemented across several other problem statements that include temporal data or require sequence modeling, and hence is a step toward the integration of XAI in temporal data. The algorithm can be improved by further analyzing ways of masking audio chunks to further explore how the absence of a signal affects the output and consequently deriving its importance. There is also the possibility of improving the speed of inference by using only an intelligently sampled subset of audio chunk combinations instead of all possible combinations. This has the potential of significantly reducing both training and inference time without compromising performance.

References

[1] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[2] R. V. Garcia, L. Wandzik, L. Grabner, and J. Krueger, "The harms of demographic bias in deep face recognition research," in *2019 International Conference on Biometrics (ICB)*. IEEE, 2019, pp. 1–6.

[3] Z. Yang, R. Al-Bahrani, A. C. Reid, S. Papanikolaou, S. R. Kalidindi, W.-k. Liao, A. Choudhary, and A. Agrawal, "Deep learning based domain knowledge integration for small datasets: Illustrative applications in materials informatics," in *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2019, pp. 1–8.

[4] S. Saravanan, K. Ramkumar, K. Adalarasu, V. Sivanandam, S. R. Kumar, S. Stalin, and R. Amirtharajan, "A systematic review of artificial intelligence (ai) based approaches for the diagnosis of parkinson's disease," *Archives of Computational Methods in Engineering*, pp. 1–15, 2022.

[5] H. Braak and E. Braak, "Pathoanatomy of parkinson's disease," *Journal of neurology*, vol. 247, no. 2, pp. II3–II10, 2000.

[6] A. Mirelman, T. Herman, S. Nicolai, A. Zijlstra, W. Zijlstra, C. Becker, L. Chiari, and J. M. Hausdorff, "Audio-biofeedback training for posture and balance in patients with parkinson's disease," *Journal of neuroengineering and rehabilitation*, vol. 8, no. 1, pp. 1–7, 2011.

[7] D. Doran, S. Schulz, and T. R. Besold, "What does explainable ai really mean? a new conceptualization of perspectives," *arXiv preprint arXiv:1710.00794*, 2017.

[8] J. Chen, M. Xie, Z. Xing, C. Chen, X. Xu, L. Zhu, and G. Li, "Object detection for graphical user interface: old fashioned or deep learning or a combination?" in *proceedings of the 28th ACM joint meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 2020, pp. 1202–1214.

[9] J. Hlavnička, R. Čmejla, Jiří Klempíř, Evžen Růžička, and J. Rusz, "Synthetic vowels of speakers with parkinson's disease and parkinsonism," 2019. [Online]. Available: https://figshare.com/articles/Synthetic_vowels_of_speakers_with_Parkinson_s_disease_and_Parkinsonism/7628819/1

[10] S. Grover, S. Bhartia, A. Yadav, K. Seeja *et al.*, "Predicting severity of parkinson's disease using deep learning," *Procedia computer science*, vol. 132, pp. 1788–1794, 2018.

[11] H. Zhang, A. Wang, D. Li, and W. Xu, "Deepvoice: a voiceprint-based mobile health framework for parkinson's disease identification," in *2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*. IEEE, 2018, pp. 214–217.

[12] O. Klempíř and R. Krupička, "Machine learning using speech utterances for parkinson disease detection," *Lékař a technika-Clinician and Technology*, vol. 48, no. 2, pp. 66–71, 2018.

[13] O. Karaman, H. Çakın, A. Alhudhaif, and K. Polat, "Robust automated parkinson disease detection based on voice signals with transfer learning," *Expert Systems with Applications*, vol. 178, p. 115013, 2021.

[14] M. Wang, W. Ge, D. Apthorp, H. Suominen *et al.*, "Robust feature engineering for parkinson disease diagnosis: new machine learning techniques," *JMIR Biomedical Engineering*, vol. 5, no. 1, p. e13611, 2020.

[15] B. Karan, K. Mahto, and S. S. Sahu, "Detection of parkinson disease using variational mode decomposition of speech signal," in *2018 International Conference on Communication and Signal Processing (ICCSP)*. IEEE, 2018, pp. 0508–0512.

[16] J. S. Almeida, P. P. Rebouças Filho, T. Carneiro, W. Wei, R. Damaševičius, R. Maskeliūnas, and V. H. C. de Albuquerque, "Detecting parkinson's disease with sustained phonation and speech signals using machine learning techniques," *Pattern Recognition Letters*, vol. 125, pp. 55–62, 2019.

[17] L. Zhang, Y. Qu, B. Jin, L. Jing, Z. Gao, Z. Liang *et al.*, "An intelligent mobile-enabled system for diagnosing parkinson disease: development and validation of a speech impairment detection system," *JMIR Medical Informatics*, vol. 8, no. 9, p. e18689, 2020.

[18] R. Fayad, M. Hajj-Hassan, G. Constantini, Z. Zarazadeh, V. Errico, A. Pisani, G. Di Lazzaro, M. Ricci, and G. Saggio, "Vocal test analysis for assessing parkinson's disease at early stage," in *2021 Sixth International Conference on Advances in Biomedical Engineering (ICABME)*. IEEE, 2021, pp. 171–174.

[19] A. Caliskan, H. Badem, A. Basturk, and M. E. Yuksel, "Diagnosis of the parkinson disease by using deep neural network classifier," *IU-Journal of Electrical & Electronics Engineering*, vol. 17, no. 2, pp. 3311–3318, 2017.

[20] M. T. Ribeiro, S. Singh, and C. Guestrin, “” why should i trust you?” explaining the predictions of any classifier,” in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[21] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadhel, M. Al-Amidie, and L. Farhan, “Review of deep learning: Concepts, cnn architectures, challenges, applications, future directions,” *Journal of big Data*, vol. 8, no. 1, pp. 1–74, 2021.

[22] R. Hyder, S. Ghaffarzadegan, Z. Feng, J. H. Hansen, and T. Hasan, “Acoustic scene classification using a cnn-supervector system trained with auditory and spectrogram image features.” in *Interspeech*, 2017, pp. 3073–3077.

[23] Y. Ho and S. Wookey, “The real-world-weight cross-entropy loss function: Modeling the costs of mislabeling,” *IEEE Access*, vol. 8, pp. 4806–4813, 2019.

[24] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, “Deepfool: a simple and accurate method to fool deep neural networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2574–2582.

[25] R. Jia and P. Liang, “Adversarial examples for evaluating reading comprehension systems,” *arXiv preprint arXiv:1707.07328*, 2017.