Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 192 (2021) 2624–2632

Procedia Computer Science

www.elsevier.com/locate/procedia

25th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems

Exploring the Impact of Data Poisoning Attacks on Machine Learning Model Reliability

Laura Verde[a,*], Fiammetta Marulli[a,*], Stefano Marrone[a]

[a]*Department of Maths and Physics, Università degli Studi della Campania "L. Vanvitelli", Caserta, Italy*

Abstract

Recent years have seen the widespread adoption of Artificial Intelligence techniques in several domains, including healthcare, justice, assisted driving and Natural Language Processing (NLP) based applications (e.g., the Fake News detection). Those mentioned are just a few examples of some domains that are particularly critical and sensitive to the reliability of the adopted machine learning systems. Therefore, several Artificial Intelligence approaches were adopted as support to realize easy and reliable solutions aimed at improving the early diagnosis, personalized treatment, remote patient monitoring and better decision-making with a consequent reduction of healthcare costs. Recent studies have shown that these techniques are venerable to attacks by adversaries at phases of artificial intelligence. Poisoned data set are the most common attack to the reliability of Artificial Intelligence approaches. Noise, for example, can have a significant impact on the overall performance of a machine learning model. This study discusses the strength of impact of noise on classification algorithms. In detail, the reliability of several machine learning techniques to distinguish correctly pathological and healthy voices by analysing poisoning data was evaluated. Voice samples selected by available database, widely used in research sector, the Saarbruecken Voice Database, were processed and analysed to evaluate the resilience and classification accuracy of these techniques. All analyses are evaluated in terms of accuracy, specificity, sensitivity, F1-score and ROC area.

Keywords: Poisoned Big Data; Data Poisoning Attacks; Security; Reliability; Resilient Machine Learning; Disorders detection; Voice quality assessment.

1. Introduction

Nowadays, the continuous connection of millions of devices led to an increase in cyber attackers, which has resulted in the need for fast and accurate detection of those attacks. Data is collected in staggering amounts these days and comes from an amazing variety of sources such as the internet, social media cell phones, multimedia applications, business archives, geolocation tools and online payments. The capability to process these great amounts of data using

* Corresponding author. Tel.: +0-000-000-0000 ; fax: +0-000-000-0000.
E-mail address: fiammetta.marulli@unicampania.it

1877-0509
10.1016/j.procs.2021.09.032

big data analytics tools in reliable and faster way can contribute to realise solutions able to detect and predictive intrusions, attacks and system vulnerabilities improving security solutions. While big data analytics is an indispensable component of any effective cybersecurity solution due to the need to process large amounts of data quickly process large amounts of data in a short time to detect possible anomalies and/or attack patterns limiting the vulnerability of such systems, on the other hand the same artificial intelligence techniques are subject to possible attacks that could compromise their reliability.

Machine and deep learning techniques are widely used in various sectors, including computer vision, natural language processing (NLP), speech recognition or healthcare anomalies detection. Many techniques built models using appropriate training data. The trained model is able to predict particular conditions, such as possible anomalies [21]. The ability to create models using training data provides hackers with the opportunity to attack learning algorithms by, for example, providing malicious inputs that could alter the efficiency and performance of the algorithm, poisoning the system. In order to have a reliable classification system, training the model with great amount of data is necessary. Having a wide data distribution, collecting training data came from several countries of the world, for example, can be useful in many applications. But opening the system to the public to provide input data means opens the system up to malicious input created by hackers to ”poison” the system. But opening up the system to the public to provide input data also equates to opening up the system to malicious input created by hackers to ”poison” the system. The episode in which numerous racist and sexist tweets were sent into Microsoft’s twitter chatbot Tay in less than 24 hours after it was opened to the public is well known [27]. Furthermore, also the rampant phenomenon of Fake News can take advantage from adopting ML and DL techniques, since it strongly relies on NLP strategies. In such a scenario, it is not uncommon to train language models to perform analysis such as stylometric. Word Embeddings Models [9, 5] are typically trained over large data set or provided as publicly available pre-trained models, by private users and/or companies, as the models provided by Facebook[1], for example. Recently, the number of attacks on NLP-based systems has significantly increased; more precisely, most of the attacks recorded in this field were Data Poisoning attacks, performed against Deep Learning and Machine Learning models and pursued by the poisoning of Word Embeddings [39, 36].

Moreover, especially in deep learning, pre-trained models are often used. However, this practice can pose a potential security risk, as publicly available pre-trained models can be attacked as backdoors. An attacker could manipulate the model to classify special inputs as a default class, while keeping the model’s performance on normal samples nearly unaffected [39]. Backdoor attacking episodes are known in computer vision area, as well as in NLP. A poisoned dataset, in which it has been added a fixed pixel perturbation or a rare word, respectively in computer vision [15] or in NLP [8], is substituted to the clean dataset altering the performances of the model.

Due to different forms of attacks, it is desirable to make these classifiers robust and resilient against such attacks, specially in healthcare sector where AI techniques are used as a valid support to early detection of possible anomalies. To improve the robustness of a classifier, one could, for example, not assign too much weight to a single feature. To distribute the weight of each feature more evenly, a regularization could be used. Several approaches were discussed in literature. A feature reweighting algorithm, useful for improving the performance and robustness of classifiers, is proposed Kolcz and Teo [19]. They also proposed a method for finding the lower bound of classifier robustness useful to evaluate each classifier. Another algorithm for reducing overweighting of each feature was proposed by Globerson and Roweis [14]. It is tested to analyse the classifiers used for handwritten digit recognition and spam filtering.

In this work, we evaluate the reliability and resilience of several machine learning (ML) techniques, able to detect voice disorders by analysing appropriate acoustic features extracted by voice samples, when data are ”poisoned”. In detail, ML approaches are capable to distinguish healthy from a voice suffering from psychogenic dysphonia. This is an alteration of the voice self-produced by the patient, unconsciously and involuntarily, in response to psychological distress. Psychogenic dysphonia is essentially characterized by the sudden or abrupt disappearance of vocal sound, by the involuntary appearance of a very breathy or whispered voice, or by an aphonic voice alternating with stretches of pressed and hyperacute voice. At the origin of the onset of symptoms, patients frequently report an inflammatory episode (real or presumed), or a negative event (illness, surgery, death of a relative), and only rarely is spontaneously linked to psychological problems; in fact, often family problems and/or work are identified [29]. Several ML technique are used to detect voice alterations. These process appropriate acoustic features to classify a voice as healthy or

[1] https://fasttext.cc/

pathological. In this study we evaluate the robustnsess not only of the main ML techniques in the case poisoned data are processed, but also of each acoustic features used as inputs of these algorithms.

The remaining sections of paper are organized as follows. Section **??** presents the main studies about the application of ML techniques to assess voice quality. The dataset, acoustic features and the analysed ML techniques are, instead, described in Section 3. Section 4 discusses the obtained results, while the conclusions are presented in Section 5.

2. Related works

Clinical voice quality evaluation is performed by using several procedures, such as the laryngeal examination, completion of self-assessment questionnaires or acoustic analysis. This consists of an estimation of appropriate acoustic parameters estimated from voice signal useful to evaluate any possible alterations of the vocal quality. Various acoustic parameters can be extracted by voice sound, such Fundamental Frequency (F_0), jitter, shimmer, Harmonic to Noise Ratio (HNR) or Mel-Frequency Cepstral Coefficients (MFCC). Appropriate tools can be used to estimate automatically these parameters, such as Multi-Dimensional Voice Program (MDVP) [2] or Praat [4] (its name deriving from the imperative form of ”praaten”, ”to speak” in Dutch), the principal systems used in clinical practice.

The acoustic features can constitute the input data of several ML algorithms able to evaluate the voice quality [35, 1]. Several ML techniques were used for voice signal processing. Among these, there are many studies in literature that identify the Support Vector Machine (SVM) as one of the main approach used to assess voice quality. The SVM algorithm was the approach used in [33] to evaluate the voice quality. MFCC constitute the input data, after to have reduct the dimensionality of data by using the Linear Discriminant Analysis (LDA), of SVM algorithm. Selvakumari et al. propose a SVM architecture to estimate possible voice alterations. Various variables like transmission energy, pitch, Silence removal, Windowing, Mel consistency and occurrence Cepstrum, and Jitter are considered [31]. The performances of SVM, Stochastic Gradient Descent (SGD) and Artificial Neural Network (ANN) classifiers are, instead, evaluate in [13]. A unified wavelet based framework is proposed to evaluate the voice quality. Energy and statistical features extracted from signals constitute the input data of these techniques.

Hidden Markov Models (HMM), Gaussian Mixture Models (GMM) and SVM are ML techniques able to classify voice samples collected in the Busan National Uni-versity Hospital by Wang et al. in [37]. As well as various machine learning techniques, that is SVM, GMM, HMM and Vector Quantization (VQ) were implemented for automatic classification of voice alterations in [23]. The Artificial Neural Network (ANN) is, instead, the technique indicated in [28] to estimate the voice quality of samples captured at the Christie and Withington Hospitals in Manchester. While the K-nearest neighbours algorithm and linear discriminant analysis is the approach proposed in [6].

Several acoustic features and ML techniques were proposed in literature for voice quality assessment. In many cases, these algorithms were trained and tested using clear datasets. In this study, we want to analyse the reliability of these techniques using poisoned data.

3. Materials and Methods

In order to evaluate the robustness of various ML techniques when data are poisoned, appropriate healthy and pathological voice samples were selected. The main acoustic parameters were exctracted, these constitutes the input data of the analysed ML algorithms.

In the following subsections, more details about the dataset, features and ML approaches were described.

3.1. Dataset

To explore the resiliency of ML techniques when data to analyse are poisoned, voice signals were selected from appropriate database, the Saarbruecken Voice Database (SVD) [25]. Voice sounds of this database were recordings at the Caritas clinic St. Theresia in Saarbruecken by the Institute of Phonetics of the University of Saarland together with the Department of Phoniatrics and Ear, Nose and Throat (ENT). It consists of more 2000 recordings of sustained /a/, /i/ and /u/ vowels and a speech sequence. Samples, freely available [26], come from subjects afflicted by several voice disorders, including functional and organic pathologies.

In this study, 91 healthy voices (mean age, 33.3 ± 17.3 years) and 91 ones come from subjects suffering from psicogenic dysphonia (mean age, 49.6 ± 10.6 years) were selected. Only adult voices were selected, to limit possible alterations and influences due to variabilities and instabilities of voice signals, typically of younger voices. Adult voices are lacking of these uncertainties that can be influence the analyses and this study. Voice quality changes, in fact, with the age. Adult voice is different from young one, due to the morphological modifications of organs that composed the pneumo-phono articulatory apparatus, responsible for voice production. This influences the voice quality providing possible alterations to our study. For this reason voices came from subjects under the age of 18 have been excluded [30, 22].

Table 1 provides more details of voice signals used in this study. The number of selected voices for each age range and gender, the percentage calculated for each group and for the complete dataset was reported.

In order to ”poison” voice samples a noise was added to sounds. The noise recording was selected by AURORA database, a noise database widely used in literature [17]. It includes noises recorded at different places, such as suburban train, crowd of people (babble), car, exhibition hall, restaurant, street, airport and train station. In this study the noises of train and speech bubble with a SNR equals to 5 dB were added to clean voice sounds that compose our dataset. in order to add noise to clean signals, Audacity tool was used. This is a free, easy-to-use, multi-track audio editor and recorder that allows the process of audio signals [3].

Table 1: Details of the voice signals used in this study.

Category	Gender	Age Group	#	%for each group	% on complete dataset
Pathological	Female	18-30	1	5.88%	0.55%
		31-50	7	41.18%	3.85%
		51+	9	52.94%	4.95%
	Male	18-30	5	6.76%	2.75%
		31-50	29	39.19%	15.93%
		51+	40	54.05%	21.98%
Healthy	Female	18-30	46	62.16%	25.27%
		31-50	12	16.22%	6.59%
		51+	16	21.62%	8.79%
	Male	18-30	12	70.59%	6.59%
		31-50	2	11.76%	1.10%
		51+	3	17.65%	1.65%
All	Female	18-30	47	51.65%	25.82%
		31-50	19	20.88%	10.44%
		51+	25	27.47%	13.74%
	Male	18-30	17	18.68%	9.34%
		31-50	31	34.07%	17.03%
		51+	43	47.25%	23.63%

3.2. Features

The voice analysis represents a very valuable technique for voice quality assessment. It consists of estimation of several parameters, able to describe objectively the characteristics of voice. The choice of feature is a fundamental task that influence the analysis and classification. In this study, the main acoustic parameters used to evaluate the voice quality were used as features for ML techniques.

In detail, the acoustic parameters evaluated are:

- Fundamental Frequency (F_0): this constitutes the rate of vibration of the vocal folds, index of laryngeal function;
- jitter: this represents the periodic variation from cycle to cycle influenced by the lack of control of vibration of the vocal folds, typical of voice disorders;
- shimmer: this relates to the amplitude variation of the voice sound and changes with the reduction of glottal resistance; and
- Harmonic to Noise Ratio: this represents the ratio of signal information over noise due to turbulent airflow, resulting from an incomplete vocal fold closure in voice alterations.

Several algorithms were used to estimate these parameters [34, 11, 32]. In this study each acoustic parameter was estimated by using Praat, a software widely used in clinical and research practice [4].

3.3. Machine Learning classifiers

In order to make an exhaustive comparison, we have chosen different ML techniques. Decision Tree is a category of ML technique used to classify categorical data in which the learned function is represented by a decision tree. Decision trees are easy to interpret,capable of working with missing values and categorical and continuous data, characteristics of the medical field. Several DT approaches were analysed:

- Random Forest: this is an ensemble learning method for classification that operates by constructing several decision trees at training time and outputting the class that is the mode of the classes. Random forests are constructed by bagging ensambles of random treed [7];
- REPTree: this algorithm builds a decision tree using the information gain and prunes it using reduced-error pruning. It is a fast decision tree learner, based on C4.5 algorithm [24];
- Random Tree: this is an ensemble supervised classifier able to generate many individual learners. A bagging approach is used to realise a random set of data for constructing a decision tree [24];
- AdaBoost: this represents the acronym for ”Adaptive Boosting” and proposed by Freund and Schapire in 1996, was the first highly successful boosting algorithm developed for binary classification. The initial classifier is constructed from the original data set. more models come generated consecutively giving more and more weight to the errors carried out in the previous models. the output of the classifier is given from the weighed sum of the predictions of the single models [10].

All the experiments were performed using the Waikato Environment for Knowledge Analysis project (WEKA), one of the most adopted framework used for classification in machine learning [12]. For each algorithm, the setting parameters are Weka’s default values. A machine with 8 GB memory and Intel(R)Core(TM) i5-6200U CPU with 2.40 GHz was adopted to perform experiments.

4. Experimental phase

The performance of the several ML techniques were evaluated in terms of accuracy, sensitivity, specificity, precision and F-score, useful to determined the ability of these algorithms to classify correctly a subject as healthy or pathological in presence or not of poisoning data.

4.1. Performance indicators of the classification

Defining the True Negatives (TNs) and Positives (TPs) the number of voice samples correctly classifies, respectively, as healthy or pathological, and False Negatives (FNs) and Positives (FPs) represent the number of samples incorrectly classifies, respectively, as healthy or pathological, the accuracy constitutes the number of correct predictions over all dataset, calculated according to equation 1:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

The sensitivity and specificity, instead, represents how many of the pathological or healthy cases the classifier correctly predicted, over all the pathological or healthy cases in the dataset. These performance metrics were estimated by using the following equations:

$$Sensitivity = \frac{TP}{TP + FN} \tag{2}$$

$$Specificity = \frac{TN}{TN + FP} \tag{3}$$

The harmonic mean of the sensitivity and precision, where the precision is the measurement of how many of the pathological predictions are correct, is calculated by the F1-score. The precision and F1-score was calculated by equations:

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

$$F1 - score = 2 * \frac{precision * sensitivity}{precision + sensitivity} \tag{5}$$

Finally, the performance of the ML techniques were evaluated considering the area under the ROC curve (AUC). This evaluates the goodness of the algorithm, when the AUC is minimum (AUC=0), the technique incorrectly classifying all samples, and when the AUC is maximum (AUC=1), the algorithm classifies perfectly healthy and pathological samples.

4.2. Results and discussion

The voice samples were divided randomly into training (80% of samples) and testing (20% of the samples) sets. We evaluated the reliability of Decision Tree techniques considering the same voice samples in two cases. In the first case we evaluated the classification accuracy on clean signals in training and testing sets. In the second case, instead, the performance of techniques were evaluated with a training set composed by clean signals, while the testing set consists of poisoned data.

Table 2: Testing results obtained considering clean signals in the training and testing sets.

Classifier	**Sensibility(%)**	**Specificity(%)**	**Accuracy(%)**	**Precision(%)**	**F1-score(%)**	**AUC**
Random Forest	83.33	83.33	83.33	83.33	83.33	0.853
Random Tree	61.11	77.78	69.44	73.33	66.67	0.694
REPTree	100.00	66.67	83.33	75.00	85.71	0.833
Adaboost	88.89	72.22	80.56	76.19	82.05	0.83

Table 3: Testing results obtained considering clean signals in the training set and poisoned signals in testing set.

Classifier	**Sensibility(%)**	**Specificity(%)**	**Accuracy(%)**	**Precision(%)**	**F1-score(%)**	**AUC**
Random Forest	100.00	61.11	80.56	72.00	83.72	0.846
Random Tree	94.44	55.56	75.00	68.00	79.07	0.819
REPTree	94.12	68.75	81.82	76.19	84.21	0.833
Adaboost	83.33	66.67	75.00	71.43	76.92	0.821

Tables 2 and 3 report the results achieved, respectively, considering clean voice samples in training and testing sets and clean samples in training set and poisoned samples in testing one. These results show,in most cases, a decrease of specificity considering a poisoned testing set compared to a testing set composed by clean signals according to a decrease of the number of true negatives, that is the number of healthy samples correctly classify as healthy. The

classifiers in presence of poisoned samples consider the healthy samples as pathological due to noise added to the signals.

In order to individuate the features more relevant to classify correctly the voice sample, and so to detect voice disorders suffering less from the effects of poisoned data, a feature selection was applied. Several feature selectors existing in literature, such as InfoGainAttributeEval [38] and Relief [20] algorithms or Principal Component Analysis (PCA) [18]. In this study a Correlation Feature Selector (CFS) was applied [16]. This estimated the predictive capability of each feature, allow to select the set of features that are highly correlated with the class and less correlated with other features, ignoring redundant and irrelevant features from the dataset. A cut-off equal to 0.20 was chosen in this study. All features that achieved a correlation rank equal or higher than this cut-off value are chosen, others were removed. Figure 1 shows the correlation rank achieved for each feature.

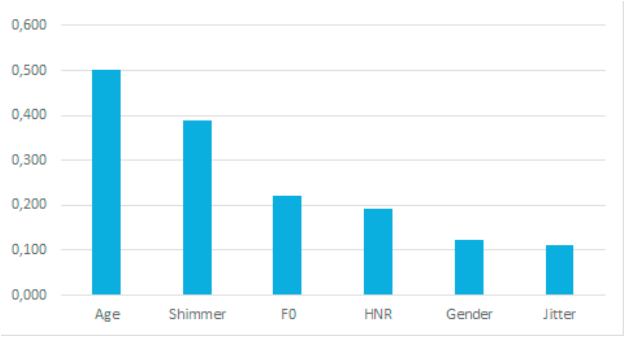

Fig. 1: Correlation rank obtained for each feature.

Considering only features that achieved a correlation rank higher than 0.20, new results obtained considering clean signals in the training set and poisoned signals in testing set are reported in table 4. These show an improvement of specificity considering the most relevant features. The best specificity value was achieved by using Random Forest (about 78%). This value is higher than the result obtained by Random Forest when all features are considered (about 61%). This demonstrates the efficiency and resilience of the selected features to represent the voice quality.

Table 4: Testing results obtained considering clean signals in the training set and poisoned signals in testing set and only parameters selected with the Correlation Feature Selector.

Classifier	Sensibility(%)	Specificity(%)	Accuracy(%)	Precision(%)	F1-score(%)	AUC
Random Forest	94.44	77.78	86.11	80.95	87.18	0.843
Random Tree	55.56	72.22	63.89	66.67	60.61	0.667
REPTree	100.00	72.22	86.11	78.26	87.80	0.889
Adaboost	94.44	66.67	80.56	73.91	82.93	0.873

5. Conclusions

Currently, Artificial Intelligence algorithms are widely adopted in several sectors, such as healthcare one. These algorithms are, in fact, use to support the early detection of particular diseases or the decision support systems to monitor patient's state of health analysing great amount of data collected from numerous sensors worn by the subjects. Unfortunately, these algorithms can be subject to attacks that may alternate their performance. This is crucial in a sector as healthcare one where the reliability and accuracy of data processing is fundamental.

In this study, we investigated the vulnerability of some Machine Learning techniques, that are Random Forest, Random Tree, REPTree and Adaboost, when poisoned data must be analysed and processed. Appropriate voices of healthy subjects and subjects suffering from psychogenic dysphonia were selected from Saarbruecken Voice Database. Acoustic features were extracted from these samples, and used as inputs of each algorithms. The aim of this study was to examine the behavior of each technique in the presence of poisoned data: The obtained results show the decrease

of specificity for each technique when poisoned data were tested. The presence of noise affects the ability of the algorithms to correctly identify healthy voices.

Our future plans provides to deep the study, analysing the expanding the dataset of voice samples as well as the number of features extracted from the signals to perform a more exhaustive investigation. Additionally, other machine learning techniques will be analysed to compare the vulnerability and reliability of several algorithms. The study of the vulnerability of the main ML models is useful to propose efficient defense methods necessary to safeguard the safety of the use of these models and their reliability in correctly supporting the diagnosis of specific diseases as well as the monitoring of the patient's health conditions.

Acknowledgements

ACK: The research described in this work is funded and realized within the activities of the the Research Program "Vanvitelli V:ALERE 2020 - WAIILD TROLS", financed by the University of Campania "L. Vanvitelli",Italy.

References

[1] Al-Dhief, F.T., Latiff, N.M.A., Malik, N.N.N.A., Salim, N.S., Baki, M.M., Albadr, M.A.A., Mohammed, M.A., 2020. A survey of voice pathology surveillance systems based on internet of things and machine learning algorithms. IEEE Access 8, 64514–64533.
[2] Amir, O., Wolf, M., Amir, N., 2007. A clinical comparison between mdvp and praat softwares: is there a difference?, in: Fifth International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications, ISCA, Firenze University Press. pp. 37–40.
[3] Audacity, T., 2013. Audacity.
[4] Boersma, P., Weenink, D., 2009. Praat: doing phonetics by computer (version 5.1. 05)[computer program]. retrieved may 1, 2009.
[5] Bojanowski, P., Grave, E., Joulin, A., Mikolov, T., 2017. Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics 5, 135–146.
[6] Boyanov, B., Hadjitodorov, S., 1997. Acoustic analysis of pathological voices. a voice analysis system for the screening of laryngeal diseases. IEEE Engineering in Medicine and Biology Magazine 16, 74–82.
[7] Breiman, L., 2001. Random forests. Machine Learning 45, 5–32.
[8] Dai, J., Chen, C., Li, Y., 2019. A backdoor attack against lstm-based text classification systems. IEEE Access 7, 138872–138878.
[9] Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 .
[10] Dietterich, T.G., 2000. Ensemble methods in machine learning, in: International workshop on multiple classifier systems, Springer. pp. 1–15.
[11] Farrús, M., Hernando, J., Ejarque, P., 2007. Jitter and shimmer measurements for speaker recognition, in: Eighth annual conference of the international speech communication association.
[12] Garner, S.R., et al., 1995. Weka: The waikato environment for knowledge analysis, in: Proceedings of the New Zealand computer science research students conference, pp. 57–64.
[13] Gidaye, G., Nirmal, J., Ezzine, K., Frikha, M., 2020. Wavelet sub-band features for voice disorder detection and classification. Multimedia Tools and Applications 79, 28499–28523.
[14] Globerson, A., Roweis, S., 2006. Nightmare at test time: robust learning by feature deletion, in: Proceedings of the 23rd international conference on Machine learning, pp. 353–360.
[15] Gu, T., Dolan-Gavitt, B., Garg, S., 2017. Badnets: Identifying vulnerabilities in the machine learning model supply chain. arXiv preprint arXiv:1708.06733 .
[16] Hall, M.A., 1999. Correlation-based feature selection for machine learning .
[17] Hirsch, H.G., Pearce, D., 2000. The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions, in: ASR2000-Automatic speech recognition: challenges for the new Millenium ISCA tutorial and research workshop (ITRW).
[18] Jolliffe, I.T., Cadima, J., 2016. Principal component analysis: a review and recent developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 374, 20150202.
[19] Kołcz, A., Teo, C.H., 2009. Feature weighting for improved classifier robustness, in: CEAS'09: sixth conference on email and anti-spam.
[20] Kononenko, I., 1994. Estimating attributes: Analysis and extensions of relief, in: European conference on machine learning, Springer. pp. 171–182.
[21] Kumar, A., Mehta, S., 2017. A survey on resilient machine learning. arXiv preprint arXiv:1707.03184 .
[22] Latoszek, B.B.v., Ulozaitė-Stanienė, N., Maryn, Y., Petrauskas, T., Uloza, V., 2019. The influence of gender and age on the acoustic voice quality index and dysphonia severity index: a normative study. Journal of Voice 33, 340–345.
[23] Mesallam, T.A., Farahat, M., Malki, K.H., Alsulaiman, M., Ali, Z., Al-Nasheri, A., Muhammad, G., 2017. Development of the arabic voice pathology database and its evaluation by using speech features and machine learning algorithms. Journal of healthcare engineering 2017.
[24] Mohamed, W.N.H.W., Salleh, M.N.M., Omar, A.H., 2012. A comparative study of reduced error pruning method in decision tree algorithms, in: 2012 IEEE International conference on control system, computing and engineering, IEEE. pp. 392–397.
[25] Pützer, M., Koreman, J., 1997. A german database of patterns of pathological vocal fold vibration. Phonus 3, 143–153.

[26] Pützer, Manfred and Koreman, Jacques, 1997. Saarbruecken Voice Database. http://www.stimmdatenbank.coli.uni-saarland.de/help_en.php4/. [Online; accessed 20-April-2021].
[27] Reuters, 2016. Microsoft's AI Twitter bot goes dark after racist, sexist tweets. https://www.reuters.com/article/us-microsoft-twitter-bot-idUSKCN0WQ2LA/. [Online; accessed 20-April-2021].
[28] Ritchings, R., McGillion, M., Moore, C., 2002. Pathological voice quality assessment using artificial neural networks. Medical engineering & physics 24, 561–564.
[29] Rosen, D.C., Shmidheiser, M.H., Sataloff, J.B., Hoffmeister, J., Sataloff, R.T., 2020. Psychogenic dysphonia. Psychology of Voice Disorders , 187.
[30] Sataloff, R.T., Linville, S., 2005. The effect of age on the voice .
[31] Selvakumari, N.S., Radha, V., 2017. A voice activity detector using svm and naïve bayes classification algorithm, in: 2017 International Conference on Signal Processing and Communication (ICSPC), IEEE. pp. 1–6.
[32] Severin, F., Bozkurt, B., Dutoit, T., 2005. Hnr extraction in voiced speech, oriented towards voice quality analysis, in: 2005 13th European Signal Processing Conference, IEEE. pp. 1–4.
[33] Souissi, N., Cherif, A., 2015. Dimensionality reduction for voice disorders identification system based on mel frequency cepstral coefficients and support vector machine, in: 2015 7th international conference on modelling, identification and control (ICMIC), IEEE. pp. 1–6.
[34] Verde, L., De Pietro, G., Sannino, G., 2018a. A methodology for voice classification based on the personalized fundamental frequency estimation. Biomedical Signal Processing and Control 42, 134–144.
[35] Verde, L., De Pietro, G., Sannino, G., 2018b. Voice disorder identification by using machine learning techniques. IEEE access 6, 16246–16255.
[36] Wallace, E., Zhao, T., Feng, S., Singh, S., 2021. Concealed data poisoning attacks on nlp models, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 139–150.
[37] Wang, J., Jo, C., 2007. Vocal folds disorder detection using pattern recognition methods, in: 2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, IEEE. pp. 3253–3256.
[38] Witten, I.H., Frank, E., Hall, M.A., Pal, C.J., 2005. Practical machine learning tools and techniques. Morgan Kaufmann , 578.
[39] Yang, W., Li, L., Zhang, Z., Ren, X., Sun, X., He, B., 2021. Be careful about poisoned word embeddings: Exploring the vulnerability of the embedding layers in nlp models. arXiv preprint arXiv:2103.15543 .