Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com/locate/artmed

Research paper

Has machine learning over-promised in healthcare? A critical analysis and a proposal for improved evaluation, with evidence from Parkinson's disease

Wenbo Ge [a,*], Christian Lueck [b,c], Hanna Suominen [a,d], Deborah Apthorp [a,e]

[a] School of Computing, Australian National University, 145 Science Road, Acton, 2601, ACT, Australia
[b] Department of Neurology, Canberra Hospital, Yamba Drive, Garran, 2605, ACT, Australia
[c] ANU Medical School, Australian National University, Hospital Rd, Garran, 2605, ACT, Australia
[d] Department of Computing, University of Turku, Turku, FI-20014, Finland
[e] School of Psychology, University of New England, Armidale, 2351, NSW, Australia

ARTICLE INFO

Keywords:
Machine learning
Parkinson's disease
Inflationary effects
Evaluation methods
Disease classification

ABSTRACT

Adoption of *artificial intelligence* (AI) by the medical community has long been anticipated, endorsed by a stream of machine learning literature showcasing AI systems that yield extraordinary performance. However, many of these systems are likely over-promising and will under-deliver in practice. One key reason is the community's failure to acknowledge and address the presence of inflationary effects in the data. These simultaneously inflate evaluation performance and prevent a model from learning the underlying task, thus severely misrepresenting how that model would perform in the real world. This paper investigated the impact of these inflationary effects on healthcare tasks, as well as how these effects can be addressed. Specifically, we defined three inflationary effects that occur in medical data sets and allow models to easily reach small training losses and prevent skillful learning. We investigated two data sets of sustained vowel phonation from participants with and without Parkinson's disease, and revealed that published models which have achieved high classification performances on these were artificially enhanced due to the inflationary effects. Our experiments showed that removing each inflationary effect corresponded with a decrease in classification accuracy, and that removing all inflationary effects reduced the evaluated performance by up to 30%. Additionally, the performance on a more realistic test set increased, suggesting that the removal of these inflationary effects enabled the model to better learn the underlying task and generalize. Source code is available at https://github.com/Wenbo-G/pd-phonation-analysis under the MIT license.

1. Introduction

Over recent years, successful implementations of *Artificial Intelligence* (AI) in the field of healthcare have become prevalent, yielding many high-performance models [1,2]. This can be seen in research surrounding the second most prevalent neurodegenerative disorder, *Parkinson's disease* (PD) [3,4], which affects approximately 1% of the population aged over 60 years [5] and is associated with impaired quality of life and an increased mortality rate [6–8]. Currently, a validated method for diagnosing and tracking the severity of PD that is free from clinicians' personal strengths and weaknesses – that is, their subject expertise – does not exist; diagnosis is made clinically on the basis of neurological history, motor examination, and response to medication [9], whilst disease tracking is often based on a questionnaire measure and a clinical motor examination [10]. This means that PD cannot easily be managed remotely and that misdiagnoses and delayed diagnoses are well recognized [11–14].

Voice recordings can be used as a minimally invasive and quantifiable marker to help address these issues, increase the level of objectivity, allow for remote management, and save clinicians' time so that they may focus more on the patient rather than on tests and data [15]. Research has shown that approximately 90% of patients with PD are affected with some kind of vocal impairment and that it may be one of the earliest indicators of the disease [16,17]. Sustained vowel phonation is similarly affected, and has the benefit of being an easier patient task that is not limited by language [4,18,19]. There already exists extensive research that utilizes *machine learning* (ML) to classify between patients with PD and healthy controls. Several results from the literature using one of the most popular data sets in this field, the

[*] Corresponding author.
E-mail address: wenbo.ge@anu.edu.au (W. Ge).

https://doi.org/10.1016/j.artmed.2023.102524
Received 18 July 2022; Received in revised form 22 February 2023; Accepted 28 February 2023
Available online 16 March 2023

Table 1
PD classification rates through analysis of sustained phonation. Accuracy is shown as percentages (% hidden for brevity). More examples can be seen in Appendix A.

	Das [23]	Li et al. [24]	Polat [25]	Zuo et al. [26]	Ma et al. [27]	Gök [28]	Ozkan [29]	Caliskan et al. [30]	Anand et al. [31]	Haq et al. [32]
Accuracy	92.90	93.47	97.93	97.47	99.49	98.46	99.10	93.80	95.00	99.00

"Parkinsons Data Set" [20,21], can be seen in Table 1. Although the results are impressive, they might not be indicative of how the models would perform in the real world [22].

We believe that much of the literature using AI for healthcare problems may be affected by artifacts that result in misleadingly high evaluation performance — so-called 'inflationary effects', which simultaneously act as a barrier to skillful learning. This gives the illusion that these models will perform well in practice when, in reality, they will not.

In this paper, we attempt to remove these inflationary effects that confound the training and evaluation process. The motive for this is so that: (1) we can assess the impact of each inflationary effect on evaluated performance, (2) models can learn the underlying task and become skilled, and (3) performance evaluations are not inflated, and instead faithfully represent the models' abilities.

2. Background

Real-world performance can be thought of as how the model would perform when tested using entirely unobserved data. In an ML context, this can also be thought of as the generalization performance; that is, the model's performance on unseen data. Here, 'unobserved' (or 'unseen') data is used to mean the set of data that has not been observed, where 'observed' refers to the data being available for training, validation, or evaluation. Their union is referred to as the population of data. For the specific task of analyzing sustained vowel phonation, the population of data encompasses the sustained vowel phonation of all people.

Clearly, it is not possible to directly assess the performance of a model on all unobserved data; thus, it must be approximated using observed data. In other words, when we evaluate a model, we are simply trying to approximate its generalization performance using observed data. Despite the simplicity of this concept, achieving a good estimate of generalization performance is often quite difficult, especially for healthcare tasks. This is sometimes referred to as the "AI chasm" [33], where the performances seen during evaluation do not translate to the real world. In an ML context, this is synonymous with the model failing to generalize.

2.1. Inflationary effects

One major reason that performances seen during evaluation do not translate to the real world is the presence of inflationary effects. These effects provide models with an avenue to easily reach a low training loss during the training phase, as models will always learn the largest and easiest classifying effect, due to the greedy nature of descent methods. Often, these inflationary effects are large and strong enough for the model to reach a near-zero training loss, which consequently halts further learning. This is not picked up as over-fitting the training set since the inflationary effects present in the training set are also present in the validation and test sets, meaning that a model which achieves low training error via inflationary effects will similarly achieve low validation and test error. This (misleadingly) suggests that the model has been trained well, will perform well on unseen data, and is likely to perform similarly well in the real world. In reality, the inflationary effects have prevented the model from skillful learning, and it will perform poorly in the real world.

We have broken down inflationary effects into three varieties. The first two seem to be fairly well-known and are occasionally accounted for in the literature, whereas the last is hardly ever acknowledged.

Digital fingerprinting phenomenon. Healthcare data often includes multiple samples derived from the same individual, which is further exacerbated when the disease, disorder, or symptom of interest has a low prevalence. Standard evaluation techniques that partition the data for training, hyper-parameter tuning, and out-of-sample testing, can result in different samples from the same individual appearing in different partitions of the data (such as both the training and test set). This may cause the model to learn to identify individuals and recall their class rather than decoding the data to gain insights into their class. This leads to a model with a low level of skill that will fail in practice, yet displays an inflated evaluation performance. The digital fingerprinting phenomenon [19,34] can occur even if an individual's identifier is removed, as their identity may be encoded within the data. Although this effect is relatively well-known, it is often not accounted for [26,31,35].

Accuracy paradox. The issue of low prevalence can also result in unbalanced class distributions, such as with classifying PD and many other healthcare tasks. Failure to account for this during evaluation can result in the model learning the skew of class distribution, ignoring the data, and blindly classifying everything as the majority class. This inflates performance yet results in a model that has no skill. The impact of the accuracy paradox [36] is well-known, and is often compensated for by the use of other performance metrics (such as the area under the ROC curve [37] or Matthew's correlation coefficient [38]). However, all performance metrics have flaws [39–41], such as being sensitive to low sample sizes, as is the case with many medical data sets. More importantly, using different metrics has no impact on model training, meaning that inflationary effects will continue to act as barriers to learning. One solution can be to weight the loss functions [42,43], which can beneficially alter the training process to handle class imbalances effectively; however, this does not handle other inflationary effects, described below.

Second order effects of the accuracy paradox. In the healthcare domain, the data used for model development (the observed data) is rarely representative of the population in regard to many factors (such as age, sex, weight, race, and so on). This imbalance of factors can be exploited by the model in a similar way to the accuracy paradox. As an example, consider the task of analyzing sustained vowel phonations to classify between controls and patients with PD. Suppose that in the observed data, the majority of patients with PD are female, whilst the majority of controls are male, a distribution that is not represented in the overall population (the true prevalence of PD leans slightly more towards males [44]). Standard evaluation techniques would result in partitions (such as the training set) that roughly mimic this distribution, from which the model can then learn the simpler task of detecting whether an individual is male or female and infer their disease state from this, ignoring the relevant information in the signal that actually relates to an individual's disease state. Such a scenario would yield a small training loss during the training phase, preventing further learning and resulting in low validation and test losses, as the same distribution would be present in both validation and test sets.

Such an effect remains regardless of class label distributions. Thus solutions to the accuracy paradox (such as using different performance metrics or altering the loss function) would do nothing to reduce the effect outlined here. As long as there are imbalanced factors and these factors are encoded into the data, this will remain an issue. Furthermore, even if factors are only slightly skewed, their combination can cause a strong enough training effect to impede the skillful learning of a model.

2.2. Generalization performance

Another reason that performances seen during evaluation do not translate to the real world is that the test sets used are often not representative of the unobserved data; that is, the test is not similar to the real world.

Typically, evaluation techniques have the underlying assumption that the observed data is representative of the population [45], and thus a random partition maintains those distributions. However, this assumption does not generally hold in a healthcare context [46]; the observed data is not always representative of the population, meaning that both the training and test sets are also not representative. Thus, what is being evaluated would not be a good approximation of generalization performance. In this case (when the test set is not representative of the population), we call this an evaluation of a model's *specific* performance to distinguish it from an evaluation of a model's *generalization* performance. That is, this is an evaluation of how well the model has learned the discriminative information under the specific guidance of this training set, from which its generalization performance cannot be inferred.

We propose that evaluating the generalization performance of a model should be done with a test set that is distributed similarly to the unobserved data, and the specific performance should be evaluated with a test set that represents the specific training set. By having two distinctly different test sets, the concern of an unskilled model accidentally being tuned to perform well on a given test set is diminished. Only a model that repeatedly shows high performance on both test sets demonstrates practical value. Further, when the dataset is augmented (such as by removing samples to match certain distributions), the specific test set often veers even further from a good generalization test set. Consider a dataset that closely matches the prevalence of PD in the population. In this case, the accuracy paradox poses a significant barrier to skillful learning. By balancing the classes, we also alter our test set such that it no longer represents the population, therefore necessitating the need to have another test set that maintains the distribution of all unseen data.

There can also be cases where all the considerations outlined above align: that is, when the distribution of the training set, the specific test set, and a good generalization test set all have the same distribution. In this case, there may not be a need to separate the specific and the general test set; however, this is unlikely to occur often in practice.

3. Methods

The code for this experiment can be found at: https://github.com/Wenbo-G/pd-phonation-analysis, under the MIT license.

3.1. Data set

Two data sets were used for this investigation: the "Parkinsons Data Set" [20,21], and the "mPower Data Set" [47]. Both consist of healthy control participants and participants with PD performing a sustained phonation of the vowel "aaahh". Summary details of both the "Parkinsons Data Sets" and the "mPower Data Set" can be seen in Table 2.

Table 2
Summary of submissions from the "Parkinsons Data Set" and filtered "mPower Data Set". Standard deviations are indicated in brackets.

	"Parkinsons Data Set"		"mPower Data Set"	
	PD group	Control group	PD group	Control group
Number of Submissions	147	48	37,047	22,148
Number of Males	97	18	19,646	18,042
Mean Age	68.0 (9.6)	60.3 (8.1)	63.9 (7.9)	40.6 (17.5)
H&Y	2.2 (0.7)	n/a	n/a	n/a
Years diagnosed	6.7 (7.1)	n/a	4.7 (5.1)	n/a

The "Parkinsons Data Set", found on the UCI Machine Learning Repository[1] [48], was recorded at the National Centre for Voice and Speech, Denver, Colorado, and was processed at the University of Oxford. Released in 2008, this data set has been widely used for detecting the presence of PD from sustained phonations and has been cited as many as approximately 900[2] times from release to the end of 2021 (according to SCOPUS). It contains 195 submissions from 31 participants, each submission consisting of 22 features extracted from the corresponding recordings. Additional data cleaning or filtering has not been performed on the provided data. The features, along with a short description, can be seen in Appendix B.1, or in the corresponding paper [21].

The "mPower Data Set" [47] was recorded via the Parkinson mPower mobile application as part of the mPower study developed by Sage Bionetworks. The first six months of the study were released in 2016, and it has been cited as many as approximately 270 times between its release and the end of 2021 (according to SCOPUS), though many of these citations are unrelated to sustained phonation, as the data set also contains many other types of data. A total of 63,000 submissions from 6,700 participants were recorded, each submission consisting of a 10-second long audio recording.

We cleaned the "mPower Data Set" by removing submissions that were either missing vital information (age, sex, or disease status), corrupted (i.e., the file could not be used), or poor in quality (i.e, did not perform sustained vowel phonation, or was contaminated with background noise). The last criterion was assessed by inspecting the "Degree of Voice Breaks (%)" and "Fraction of Locally Unvoiced Frames (%)" extracted using Praat [49], along with root-mean-square, variability, and energy of the audio signal (more details can be found in Appendix B.2). Following this, the same 22 features were extracted from the valid recordings using the same protocol as the "Parkinsons Data Set" [20]. Prior to feature extraction, we removed the first 2 s and final 3 s (leaving a 5-second window between the 2 and 7-second mark), to eliminate artifacts caused by the participant starting the task late or running out of breath. Finally, we manually inspected samples that produced the 20 most extreme values (10 highest and 10 lowest) for each feature, repeated for all features, to ensure that the assessment of quality was good.

3.2. Evaluating specific performance and the impact of inflationary effects

To determine the impact of the inflationary effects, several successful models from the literature were chosen to be replicated on both data sets. First, these models were evaluated on the data without any modification, to assess the performance when all inflationary effects

[1] https://archive.ics.uci.edu/ml/datasets/parkinsons

[2] This is the number of citations, excluding duplicates, of both papers [20, 21] that were requested to be cited when using the data set.

were present. Each inflationary effect was then sequentially removed to show its marginal contribution across the different models and data sets.

We looked to The Association of Computing Machinery's definition of "replicate" – to be able to obtain the same results when an independent group uses the original author's own artifacts (that is, "different team, same experimental setup") [50]. Three models from the literature were chosen: the first achieved an accuracy of 99.1% by reducing the input space with *Principal Component Analysis* (PCA) and classifying with the *k-Nearest Neighbor* algorithm (*k*-NN) [29]. The second method achieved an accuracy of 93.79% and classified participants with a *stacked AutoEncoder* (sAE), which progressively encodes the features into a restricted latent space to be used for classification [30]. The third method achieved an accuracy of 99% by using separate *Support Vector Machines* (SVM), one with an L1-loss for embedded feature selection, and another with an RBF kernel to classify the set of restricted features [32]. These are referred to in this text as **Model A**, **Model B**, and **Model C**, respectively. The evaluation method used in the original papers, repeated 10-fold *cross-validation* (CV) and repeated 70/30 train/test splits [51], was used here, with 30 repetitions for the "Parkinsons Data Set", and 10 repetitions for the "mPower Data Set". Grid search was also used to find a good set of hyper-parameters for each model using the same evaluation methods. All repetitions were similarly seeded within a given repetition. Additional details of the grid search can be seen in Appendix B.4.

The marginal impact of each inflationary effect was shown in an ablation fashion. First, the models were evaluated on the data sets without change, representing a baseline of performance where all inflationary effects were present. Next, the digital fingerprinting phenomenon was removed by ensuring that there was no participant cross-over between each data partition across both the *k*-fold CV and train-test split evaluation methods. Following this, the impact of the accuracy paradox was eliminated by removing samples until there was the same number of samples from each class in all subsets of the data, thus balancing the class distribution. Finally, the second-order effects of the accuracy paradox were removed by ensuring that the age distribution for each combination of sex and class was similar in all subsets of the data. We did so because the distribution of age and sex in the data set was imbalanced, and because the age and sex of an individual are likely to be encoded in their voice/sustained phonation, therefore making these factors a possible contributor to the inflationary effects of the second order accuracy paradox. This was accomplished by applying a method used in Wang et al. [19] to find a "twin" of the opposite class for every sample: for every submission from a patient with PD of a given sex and age, a "twin" submission from the control group was found with the same sex and similar age (± 3 years). If a "twin" could not be found, that sample was discarded. All pairs were then partitioned using the standard evaluation methods mentioned previously, resulting in subsets of data for which the joint distribution of age-sex was similar. An underlying assumption of this method is that if every single person had an exact twin, the only difference being the presence or absence of a disease in the twin, using such a data set would result in the most skillful learning. Since this is an impossibility, the next best thing would be to use "digital twins"; that is, similar in many factors but different in disease status.

The pseudo-code for the above can be found in Appendix D. The key benefits of this method are its robustness to the prior aforementioned inflationary effects, as well as the potential to handle other inflationary factors.

The combination of the aforementioned constraints meant that the original 10-fold CV was no longer viable due to the difficulty in identifying 10 subsets that met all requirements. Therefore, we used an 8-fold CV for the "Parkinsons Data Set", and a 5-fold CV for the "mPower Data Set". No changes were made to the number of repetitions. Additionally, age limitations for the smaller "Parkinsons Data Set" were relaxed for the female participants, and these participants were instead paired by hand to their best age match.

In order to determine if any changes in evaluated performance were attributed to the reduction in training samples as opposed to the removal of inflationary effects, we performed the same baseline experiments, but randomly discarded samples until the number of remaining samples was equal to if all inflationary effects were removed. This maintained the inflationary effects but corresponded to the same reduction in the number of samples.

3.3. Generalization performance

The generalization performance (or real-world performance) was approximated by combining both the "Parkinsons Data Set" and the "mPower Data Set", then setting aside a test set that was representative of the unobserved data; that is, data the model would see in the real world. This meant that in the test set, each participant only had one sample, the prevalence of PD was low, half the controls were male, two-thirds of the patients with PD were male, and PD was more prevalent in the older participants. The remainder of the combined data set was used for training and hyper-parameter optimization through 5 times repeated 5-fold CV and grid search (details of the search space can be found in Appendix B.4). The entire process (sampling the test set, removing the inflationary effects, and the grid search) was repeated 10 times and averaged.

Specifically, the test set consisted of 100 participants: 6 males with PD, 4 females with PD, 45 males without PD, and 45 females without PD. The square root of age was used as weighting when sampling healthy controls, and the square of age was used as weighting when sampling patients with PD. This was to create a test set in which older participants were more likely to have PD, and younger participants were more likely to be healthy controls (similar to the real world). Note that both weighting methods translated to older participants being picked more often, but this was less exaggerated for the controls. Finally, at least 2 controls and 4 patients with PD were taken from the "Parkinsons Data Set", due to the size difference between the two data sets.

Although the prevalence of PD here remained higher than that of the real world (approximately 1% of people over 60), a trade-off was needed due to data limitations. A realistic prevalence would provide a better approximation of generalization performance, but would necessitate many more samples to allow the model to express its ability in detecting PD; a test set with 1 PD participant and 99 controls would not give the model much chance to show its capabilities, whilst a test set with 10 PD participants and 990 controls would leave little for the training set. For this reason, we compromised with a 10% prevalence rate (10 PD participants and 90 controls).

In addition to this, the same experiment was repeated with inflationary effects present. That is, the repetition made no modifications to the remaining data after the test set had been set aside, to evaluate the difference in generalization performance.

4. Results and analysis

4.1. Evaluating specific performance and the impact of inflationary effects

The removal of each inflationary effect resulted in a corresponding decrease in performance, across all models, both evaluation methods, and both data sets. The removal of all inflationary effects yielded an average decrease in accuracy of approximately 29.7% for the

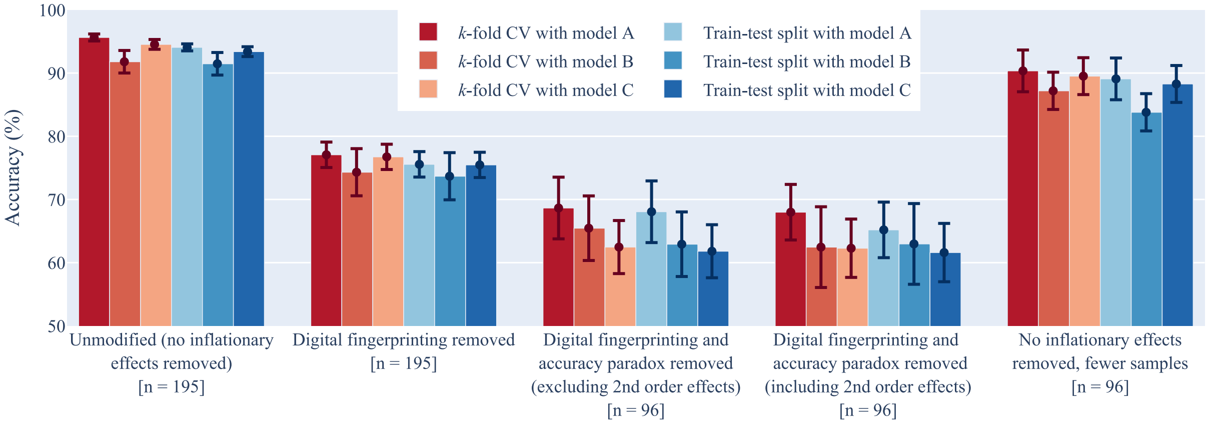

Fig. 1. The marginal impact of each inflationary effect with the “Parkinsons Data Set”.

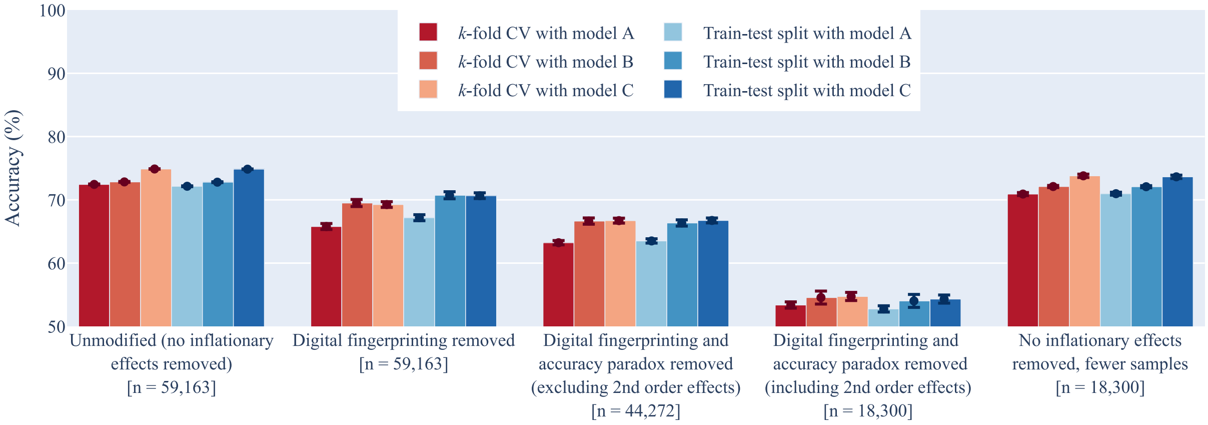

Fig. 2. The marginal impact of each inflationary effect with the “mPower Data Set”.

“Parkinsons Data Set”, and 19.4% for the “mPower Data Set”, shown in Figs. 1 and 2, and in Tables 3 and 4. Notably, the marginal impact of the second-order effects of the accuracy paradox seemed to be much greater with the larger data set. Similar results were observed for other performance metrics, such as AUROC and *Matthew’s correlation coefficient* (MCC) [38], and can be seen in Appendix C.2. It should also be noted that the removal of the second-order effects of the accuracy paradox in the “Parkinsons Data Set” (Fig. 1) did not reduce the number of samples but simply rearranged them, as a match could be found for each participant.

Reducing the number of training samples whilst preserving the inflationary effects indicated that the decrease in performance could be mainly attributed to the removal of inflationary effects. This can be seen by comparing the first cluster with the two right-most clusters in Figs. 1 and 2. Although removing training samples without removing inflationary effects resulted in a small performance decrease (an average of approximately 5.4% and 1.1% for the “Parkinsons Data Set” and the “mPower Data Set”, respectively), this was minor compared to removing the inflationary effects . More detailed results can be seen in Appendix C.2.

4.2. Evaluating generalization performance

An approximation of the generalization performance of each model can be seen in Table 5. The poor generalization performance, as well as the low specific performance shown in the prior section, suggests that the models and methods from the literature would perform poorly in the real world. It is not clear whether this is due to a lack of information in the data or a lack of capacity in the models, but it is clear that despite the overwhelming amount of literature that shows high performances for this task, much more research is required.

Table 3
Average accuracy and AUROC of repeated k-fold cross validation on the "Parkinsons Data Set" with and without the presence of inflationary effects, for models A, B, and C. Accuracy and AUROC are shown as percentages (% hidden for brevity), and standard deviations are shown in brackets.

Inflationary Effects	Model	Accuracy	AUROC
Present	A	95.6 (0.6)	95.4 (1.1)
	B	91.8 (1.8)	96.7 (1.4)
	C	94.5 (0.8)	98.2 (0.7)
Removed	A	68.0 (4.4)	67.8 (4.4)
	B	62.5 (6.4)	70.3 (8.8)
	C	62.3 (4.6)	69.0 (7.9)

Table 4
Average accuracy and AUROC of repeated k-fold cross validation on the "mPower Data Set" with and without the presence of inflationary effects, for models A, B, and C. Accuracy and AUROC are shown as percentages (% hidden for brevity), and standard deviations are shown in brackets.

Inflationary Effects	Model	Accuracy	AUROC
Present	A	72.4 (0.1)	77.6 (0.1)
	B	72.8 (0.1)	78.4 (0.1)
	C	74.9 (0.0)	80.8 (0.0)
Removed	A	53.4 (0.5)	54.7 (0.7)
	B	54.6 (1.0)	57.3 (1.1)
	C	54.7 (0.6)	56.9 (0.9)

Table 5
Generalization performance evaluated with a generalization test set. Accuracy is shown as percentages (% hidden for brevity), and standard deviations are shown in brackets. n is the total number of samples used.

Changes to the training set	Model	Accuracy
Inflationary effects present [$n = 59,258$]	A	42.6 (5.3)
	B	45.8 (6.1)
	C	45.4 (4.9)
Inflationary effects removed [$n = 18,395$]	A	52.3 (4.3)
	B	53.8 (8.0)
	C	58.8 (5.3)
Fewer samples, inflationary effects present [$n = 18,395$]	A	42.4 (3.9)
	B	43.2 (3.8)
	C	42.7 (4.2)

The same process was repeated without removing any samples, as well as randomly removing samples until the number of remaining samples was equal to the number of samples in a data set with the inflationary effects removed, both maintaining the inflationary effects. This revealed that the generalization performance was indeed worse in the presence of the inflationary effects and that the reduced sample size did not contribute to the improved performance when eliminating inflationary effects. This suggests that these inflationary effects prevent skillful learning which would generalize better in a practical environment. However, this trend was not seen with other performance metrics, shown in Appendix C.3.

4.3. Limitations

The findings outlined here are subject to some limitations. The original "Parkinsons Data Set" was recorded with specialized equipment in a supervised and controlled environment, whereas the "mPower Data Set" was recorded on a smartphone in an uncontrolled environment, which may have affected the quality of the recordings. Furthermore, the data cleaning step for both the "Parkinsons Data Set" and "mPower Data Set" could not be matched exactly due to different data types and missing details. Instead, the data cleaning step for the "mPower Data Set" attempted to filter out the recordings of low quality (such as recordings that were silent, corrupted with background noise, or incomplete due to interruptions). Despite this, several low-quality recordings may still remain, though this should not have a strong enough learning effect to impact the model significantly [52]. The feature extraction method was also slightly different; the correlation dimension (D2) feature could not be extracted with the original software and was instead implemented in Python following the same algorithm and practices [53,54]. Further details of this can be found in Appendix B.3. Additionally, there may be slight differences in the feature extraction implementations for both data sets. However, this should not have a significant impact when the data sets are used independently and should only have minor effects when combined due to the differences in size.

5. Discussion

This study revealed the presence of inflationary effects and their ability to produce a model that appears to perform well when evaluated with standard machine learning methods but, in reality, has failed to train skillfully and will perform poorly in the real world. Broadly, this falls under the category of reproducibility (or lack thereof) in AI [55]. By investigating the classification of PD patients from healthy controls using sustained vowel phonation and by modifying the data to eliminate each individual inflationary effect, we found that the removal of each inflationary effect corresponded with a decrease in performance and that the removal of all inflationary effects resulted in a decrease of up to 30% accuracy. Furthermore, when evaluating with a test set that was constructed to be similar to the real world, the absence of inflationary effects resulted in an increased performance compared to when inflationary effects were present, suggesting that corresponding barriers to learning were also removed, allowing for the model to train better and allowing for a more accurate evaluation of generalization performance, which in turn informs practical value.

Emerging in the literature are experiments that document similar issues to those we have outlined. Ozbolt et al. [56] uses the "mPower Data Set" and several other data sets to outline considerations when detecting PD using phonation of sustained vowels. These issues include record-wise and subject-wise folds (which we termed the digital fingerprinting phenomenon) and the age imbalance within each class (which we have generalized as the second-order effects of the accuracy paradox). Our research differs from this by outlining the general framework for second-order effects of the accuracy paradox, enabling us to consider and account for the joint distribution of age and sex across the classes. This effect of age and sex on speech performance in PD is further supported by two recent phenotypic studies that have shown the significant effect of these two factors [57,58].

The contributions outlined in this article aim to bring awareness to some of the barriers that prevent skillful training of models in a healthcare context, and also to gain a better estimate of their generalization performance. We hope this will result in a reduction in the number of publications that proclaim over-optimistic results in the literature. There already exists a plethora of experiments that use the same or similar data sets and seem not to have taken into account inflationary effects [59–62]. These exude a false sense of confidence in the real-world performance that the models can offer. Even when some inflationary effects are appropriately managed (such as the more commonly known accuracy paradox or digital fingerprinting effect), the remaining effects still impact the training and evaluation of the model, resulting in misleading conclusions. In extreme cases, this can cause the premature adoption of AI models, which can have widespread effects and severely impact the health of many patients. Using IBM's Watson for Oncology [63] as an example, this model was trained with a small number of 'synthetic' cancer cases and often gave multiple unsafe and incorrect treatment recommendations, verified by medical specialists and customers [64,65]. Having a more accurate estimate of the model's general ability may have prevented this, which would not only prevent potential harm to patients but also halt the degradation of the relationship between academia and practice. We must be mindful that patients come first; nothing should override the stringent requirements for clinical deployment [66].

In addition to the always-present goal of achieving better performance, future works surrounding the classification of PD should aim to uncover additional factors which may inflate performance through the second-order effects of the accuracy paradox, as well as determine how the performance of these models extrapolate to recently diagnosed PD patients [67]. We must ensure that our evaluation methods are robust, and we must also remember that there is little clinical use in correctly classifying PD patients after they have been diagnosed. Once models show high performance after robust evaluation, the adoption for their applications (such as for early diagnosis or differential diagnosis) can be supported through the use of randomized controlled trials, a technique that is under-utilized within AI for healthcare [68].

6. Broader impacts

These issues also more broadly impact aspects of society, such as fairness and equality, as the inflationary effects can be present whenever data are derived from humans [69,70]. An example can be seen in policing algorithms assessing the recidivism risk (the risk that a convicted criminal will re-offend) [71]. If the observed data contains a high proportion of one race in the 'will re-offend' class, and a high proportion of another race in the 'will not re-offend' class, standard evaluation procedures will result in the model learning to unfairly classify based on race rather than insights within the data. This becomes more significant when race is not directly given and is encoded in the data (possibly through the location of arrest, an individual's address, or level of education), as it may give practitioners false confidence that the model is race-agnostic when it is not. This further extends to minority groupings not defined by race. Such hidden inflationary effects also become an issue in data-imputation tasks [72] as biases encoded into the model will likely flow through to the imputation step, biasing the data even further, compounding the issue. By utilizing the "twin" pairing method [19], along with a separate test set used to evaluate generalization performance that is representative of the population, fair treatment is ensured by eliminating the inflationary effects, and performance in the real world can be more accurately estimated.

In discussing the positive outcomes, we must also discuss the potential harms our methodology may cause. One of these stems from the difficulty of dealing with the second-order effects of the accuracy paradox, specifically due to not knowing beforehand which factors are encoded in the data, exacerbated by a large number of potential factors. This becomes increasingly difficult if the encoded factors are not part of the data set. By eliminating the impact of a few factors, we may have trained the model to learn the underlying task better, but we may also be forcing the model to rely more heavily on other factors (such as sex or race). Another potential harm is the possibly biased construction of the generalization test set. By choosing the data to be representative of the population in several factors, we may introduce our own biases and corrupt the purpose of fair evaluation. If a model showed high specific and general performance under these errors, it might provide enough false confidence to put into production an unfair model, potentially causing widespread harm.

To minimize the downsides, we suggest future research in systematic methods to identify the at-risk factors that may result in the second-order effects of the accuracy paradox (such as race, sex, handedness, age, or body-fat percentage), what can be done when these at-risk factors are not known (for example, if the age of the participant was not given in either data sets used here), and alternative methods for addressing the inflationary effects (as opposed to the digital "twin" pairing method).

We believe that medical data sets sit in a unique position that can inform further ethical concerns for AI in general. Unlike many human-derived data sets for non-medical tasks, medical data sets often contain additional information about the participants (such as weight, age, and pre-existing conditions). This allows for the investigation of how to control exploitable factors best when they are known, which can then lead to the research of how to control these factors when they are not known. Such developments are paramount to fairness and equality in AI. Consequently, we propose supplementing relevant guiding principles and policies (such as the U.S. FDA's Good Machine Learning Practice for Medical Device Development: Guiding Principles [73]) with the findings of this investigation.

We also stress the importance of collaboration between ML experts who develop the models, medical experts who use the models, and those who gather the data. This joint effort up front would generate a frugal and purposeful approach to creating research data [74]; minimal data wastage would result because target distribution built with the intention of having minimal inflationary effects would have been designed before gathering any data.

To conclude, our experiments indicated that inflationary effects in AI healthcare tasks can prevent skillful learning by a model which may yet show good performance. Our investigations also showed that these effects could be mitigated by removing barriers that prevent the model from learning the underlying task. We also present an alternative method to evaluate a model's generalization, thereby gaining a better estimate of its practical value and skill. Applying these ideas to the task of classifying PD patients through their sustained phonation showed that an overwhelming amount of literature shows inflated performances, and the reality is that further research is required before machine learning can be helpful in practice.

Ethics statement

Our experiments included data from human participants, and consequently, we obtained the proper ethics approvals and research permissions. The use of these existing data sets in our study was covered under the ANU Human Ethics Protocol number 2018/108 and certified through the Synapse Awareness and Ethics Pledge. No new data was collected as part of this study.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We gratefully acknowledge funding from the Australian Government Research Training Program Domestic Scholarship for the first author's PhD studies. This research has been delivered in partnership with Our Health in Our Hands (OHIOH), a strategic initiative of the Australian National University, which aims to transform healthcare by developing new personalized health technologies and solutions in collaboration with patients, clinicians, and healthcare providers. Additionally, we thank the research participants and study investigators of the mPower study. These data were contributed by users of the Parkinson mPower mobile app, as part of the mPower study developed by Sage Bionetworks, described in Synapse [doi:10.7303/syn4993293].

Finally, we would like to thank Prof. Max A. Little for his guidance in the feature extraction process; the study would be incomplete without it.

Appendix A. Experimental performances from literature

Here we show the results from several experiments in literature for the same task, that is, distinguishing if a recording of sustained vowel phonation belongs to a patient with PD or a control. Many of these have used the same "Parkinsons Data Set", but several have used other data sets consisting of similar features. This is not a comprehensive list, but rather just the results from a simple search of papers that cite the "Parkinsons Data Set" and a quick scan to determine if the same task was performed. Table A.6 lists several papers, with references.

The following Fig. A.3 shows the number of times the "Parkinsons Data Set" and the "mPower Data Set" has been cited since it was released, according to SCOPUS, as of June 2022. There are two lines for the "Parkinsons Data Set" ("Little et al. 2009" and "Little et al. 2007") as the authors have requested both papers to be cited when using the data set. The purpose of illustrating this trend is to highlight the fact that research in this field is increasing, and without proper evaluation methods, inflated performances will continue to flood the literature.

Table A.6
Performance of several papers in the literature performing the same task of distinguishing if a sustained vowel phonation belongs to a person with PD or a control. **Bold** indicates the best 3 performances, whilst *italics* indicates to worst 3 performances.

Authors	Year	Accuracy
Das [23]	2010	92.9
Sakar and Kursun [75]	2010	92.8
Li et al. [24]	2011	93.5
Ozcift and Gulten [76]	2011	*84.4*
Polat [25]	2012	97.9
Sakar et al. [77]	2013	*85.0*
Zuo et al. [26]	2013	97.5
Ma et al. [27]	2014	**99.5**
Gok [28]	2015	98.5
Tang and He [78]	2015	*77.7*
Abiyev and Abizade [79]	2016	**100.0**
Ozkan [29]	2016	99.1
Caliskan et al. [30]	2017	93.8
Guruler [80]	2017	**99.5**
Cai et al. [59]	2018	97.0
Aich et al. [81]	2018	96.9
Ul Haq et al. [32]	2019	99.0
Anand et al. [31]	2019	95.0

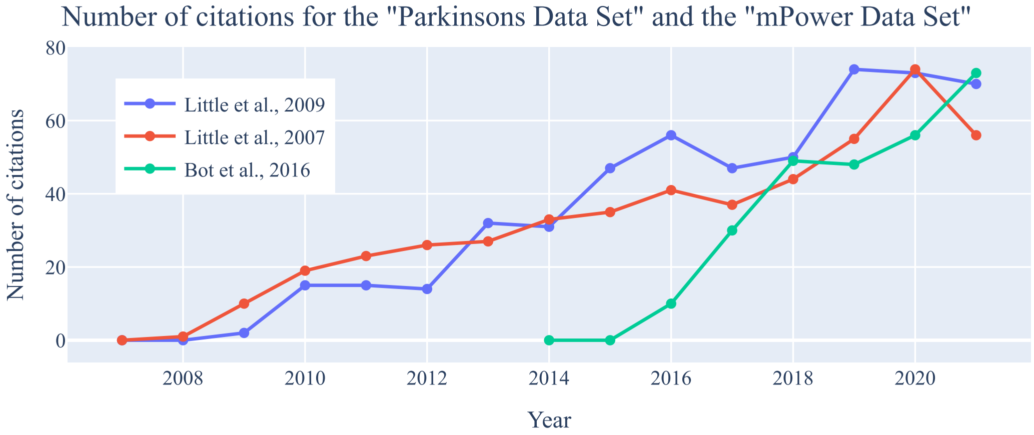

Fig. A.3. Number of citations for the "Parkinsons Data Set" and the "mPower Data Set".

Appendix B. Additional parameters in the methods

This section of the appendix outlines several parameters in the methods that were not explicitly stated in the main text for brevity. This is for the purposes of reproducibility, and contains the values used for filtering recordings, the parameters used to extract the correlation dimension, and the search space used for the grid search. In addition to these, the code can be found on https://github.com/Wenbo-G/pd-phonation-analysis. This repository will be released under the MIT license.

B.1. Feature names and description

The names of the features used, as well as the descriptions, can be found in Table B.7. This can also be found in the online repository for the data (https://archive.ics.uci.edu/ml/datasets/parkinsons), or in the corresponding papers [20,21].

Table B.7
Feature names and descriptions.

Feature name	Description
MDVP:Fo(Hz)	Average vocal fundamental frequency
MDVP:Fhi(Hz)	Maximum vocal fundamental frequency
MDVP:Flo(Hz)	Minimum vocal fundamental frequency
MDVP:Jitter(%), MDVP:Jitter(Abs), MDVP:RAP, MDVP:PPQ, and Jitter:DDP	Several measures of variation in fundamental frequency
MDVP:Shimmer, MDVP:Shimmer(dB), Shimmer:APQ3, Shimmer:APQ5, MDVP:APQ, and Shimmer:DDA	Several measures of variation in amplitude
Noise-to-harmonics ratio and Harmonics-to-noise ratio	Two measures of ratio of noise to tonal components
RPDE and D2	Two nonlinear dynamical complexity measures
DFA	Signal fractal scaling exponent
spread1, spread2, and PPE	Three nonlinear measures of fundamental frequency variation

B.2. Threshold values for filtering recordings

The sustained phonation recordings were filtered to remove those of low quality. This was assessed through several features: “Degree of Voice Breaks (%)” and “Fraction of Locally Unvoiced Frames (%)” (both extracted from Praat software [49]), as well as the RMS of amplitude for 1 s intervals, and energy (the integral of the signal squared) for 1 s intervals, between the 2 s and 7 s mark of the audio recording. The thresholds for each of the features can be seen in Table B.8.

Table B.8
Features and threshold values to filter recordings. The condition and value define a recording that passes the filter, that is, not a bad recording.

Feature	Condition	Value
Mean of RMS	>	300
Standard deviation of RMS	<	2,000
Mean of energy	>	50,000
Degree of voice breaks (%)	<	30
Fraction of locally unvoiced frames (%)	<	30

B.3. Parameters for extracting the correlation dimension

Since the feature D2 (Correlation Dimension) could not be extracted with the same software (TISEAN [53]), we implemented the same algorithm (the Grassberger and Procaccia algorithm, adapted by Theiler [82,83]) in python. The hyper-parameters were determined through various tests: the time delay (τ) determined through mutual information [84], embedding dimension (m) determined with false nearest neighbors [85], and the Theiler window (w) was determined with the space time plot [86], as recommended in the original software [53]. The exact values used for these can be seen in Table B.9.

Table B.9
Parameter values used to extract the Correlation Dimension (D2) from the recordings in the “mPower Data Set”.

Parameter	Value
Time delay (τ)	18
Embedding dimension (m)	10
Theiler window (w)	30

B.4. Grid search details

This section outlines the search space of the grid search used for all the models (A, B, and C) for each data set (“Parkinsons Data Set” and “mPower Data Set”), and for each of the different purposes (evaluating the skill and evaluating the generalization performance), shown in Tables B.10 and B.11. The search space was designed to capture the original best model outlined in the original experiments, as well as searching a few additional parameters that may be beneficial to explore.

Note that the search space for the generalization performance was much smaller than the search space when assessing the skill of the model. This is because it was partially informed by the hyper-parameters that performed well during skill evaluation, but also because the goal was not to find the best-performing model, but rather to compare the differences in performance with the removal of the various inflationary effects. Also recall that the data sets were combined when evaluating the generalization performance, so it was not split into the corresponding data sets.

Table B.10
Hyper-parameter search space for Models A, B, and C, for evaluating skill. Bold are the hyper-parameters that resulted in the best performance.

	Model	Hyper-parameter	Search space
"Parkinsons Data Set"	A	Number of principle components	2, 5, 8, 11, **14**
		Number of neighbors (k)	**1**, 3, 5, 7, 9, 11
		Preprocessing scaler	Standard, MinMax
	B	Learning rate	**0.003**, 0.03
		Number of epochs	100, 200, **400**
		Final activation in AutoEncoder	Tanh, Sigmoid, ReLU
		Latent size	4, **6**
		Preprocessing scaler	Standard, MinMax
	C	Kernel	Radial basis function, linear
		Gamma	'scale', 'auto', 0.04, 0.075, 0.09, **0.2**, 0.4
		C	1, **5**, 10
		Number of features	6, 8, 10, 12, **14**, 16, 18, 20, 22
		Preprocessing scaler	Standard, MinMax
"mPower Data Set"	A	Number of principle components	5, 8, 12, **16**
		Number of neighbors (k)	1, 5, 9, **11**
		Preprocessing scaler	Standard, MinMax
	B	Learning rate	**0.003**, 0.03
		Number of epochs	10, **50**
		Final activation in AutoEncoder	Sigmoid, ReLU
		Latent size	4, **6**
		Preprocessing scaler	Standard, MinMax
	C	Kernel	Radial basis function, linear
		Gamma	'scale', 'auto', 0.0001, 0.005, 0.1, 0.2
		C	1, **10**
		Number of features	5, 10, 15, **20**
		Preprocessing scaler	Standard, MinMax

Table B.11
Hyper-parameter search space for Models A, B, and C, for generalization performance. Note that there was no best set of hyper-parameters as they always changed between runs (due to random sampling).

	Model	Hyper-parameter	Search space
"Parkinsons Data Set"	A	Number of principle components	8, 12, 16
		Number of neighbors (k)	5, 9, 11
		Preprocessing scaler	Standard
	B	Learning rate	0.003, 0.03
		Number of epochs	50
		Final activation in AutoEncoder	Sigmoid, ReLU
		Latent size	4, 6
		Preprocessing scaler	Standard
	C	Kernel	Radial basis function
		Gamma	'scale', 'auto', 0.2
		C	1, 10
		Number of features	10, 15, 20
		Preprocessing scaler	Standard

B.5. *Female participant matching in "Parkinsons Data Set"*

Due to the small number of females in the "Parkinsons Data Set", the age limitations were relaxed, and the "twins" were matched by hand for the best pairing. This is shown in Table B.12. Each row is a pair.

Table B.12
Matching PD and control females for the "Parkinsons Data Set". Each row is a pair.

PD participant		Control participant	
Participant code	Age	Participant code	Age
S08	48	S10	46
S26	53	S07	48
S06	63	S17	64
S05	72	S42 or S50	66 or 66
S34 or S21	79 or 81	S50 or S42	66 or 66

Appendix C. Additional results

This section of the appendix outlines several additional results not presented in the main text. It contains more details about the data sets used, the performances resulting from the removal of each inflationary effect, the results of restricting the size of the data sets, and the generalization performances when training with an unmodified data set. All performances shown here have additional metrics not shown in the main text. If more details are desired, the confusion matrices (that is, the number of true positives, false positives, true negatives, and false negatives) can be found in the open-source repo: https://github.com/Wenbo-G/pd-phonation-analysis. The repository will be released under the MIT license.

C.1. Demographics of the “Parkinsons Data Set” and “mPower Data Set” after inflationary effects were removed

Table C.13 shows a summary of the demographics of both data sets after all the inflationary effects have been removed. Note that the resultant data set is not always the same as the pairs are chosen randomly. The information shown here is just one instance of the resulting data.

Table C.13
Demographics of submissions in the “Parkinsons Data Set” and “mPower Data Set” after inflationary effects have been removed. Standard deviation indicated in brackets.

	“Parkinsons Data Set”		“mPower Data Set”	
	PD group	Control group	PD group	Control group
# Submissions	48	48	9161	9161
# Males	18	18	7559	7559
Mean Age	63.25 (9.37)	60.25 (8.10)	59.04 (10.33)	58.32 (11.26)
H&Y	2.44 (0.42)	n/a	n/a	n/a
Years diagnosed	7.34 (9.53)	n/a	4.52 (4.49)	n/a

Additionally, the distribution of the age across the sex and class for the unmodified data set and the modified data set (that is, inflationary effects removed) can be seen in Figs. C.4 and C.5. Note that this shows the distribution of the samples, not the distribution of participants.

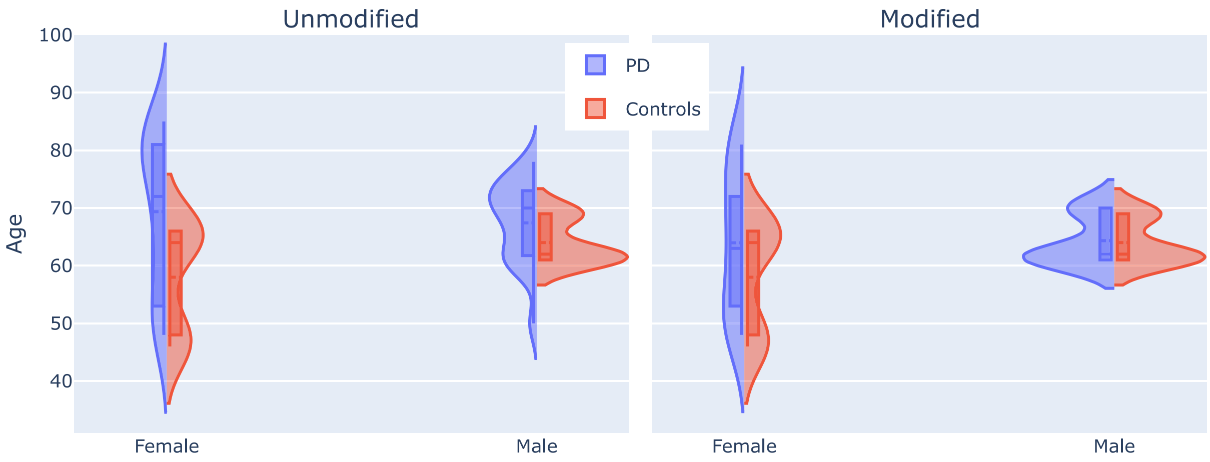

Fig. C.4. Age distribution across class and sex for the “Parkinsons Data Set”.

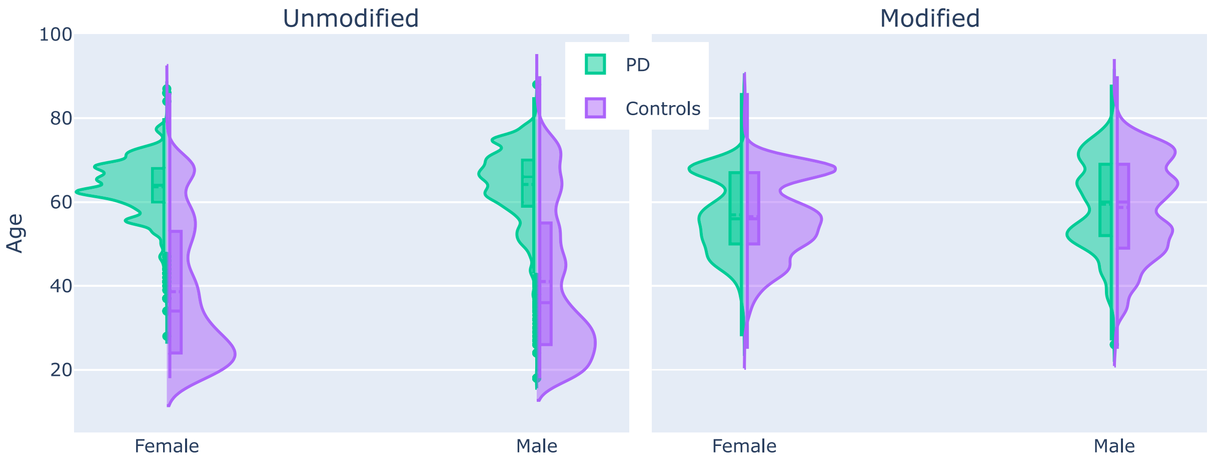

Fig. C.5. Age distribution across class and sex for the “mPower Data Set”.

C.2. Full performance details for experiments

The full results from the removal of each inflationary effect can be seen in Tables C.14 and C.15, for both data sets. The results show that in almost all cases, the removal of each inflationary effect is accompanied with a decrease in performance. It also shows that reduction in the number of samples available is not a significant contributor to the reduction in performance. We can see that the decrease in available samples decrease the performance by an average of 5.39% and 1.06% for the “Parkinsons Data Set” and the “mPower Data Set”, respectively. In contrast to this, the removal of the inflationary effects resulted in an average decrease in performance of 29.76% and 19.21% for the “Parkinsons Data Set” and the “mPower Data Set”, respectively.

Table C.14
Full results with models A, B, and C with the "Parkinsons Data Set". Accuracy and AUROC are percentages (% omitted for brevity). Standard deviations shown in brackets.

	Model	Evaluation method	Accuracy	MCC	AUROC
Unmodified (no inflationary effects removed)	A	repeated 10-fold CV	95.63 (0.55)	0.89 (0.02)	95.40 (1.12)
		repeated train-test split	94.07 (2.77)	0.85 (0.07)	93.25 (3.81)
	B	repeated 10-fold CV	91.79 (1.78)	0.79 (0.05)	96.75 (1.44)
		repeated train-test split	91.47 (3.40)	0.78 (0.09)	95.60 (3.49)
	C	repeated 10-fold CV	94.54 (0.78)	0.85 (0.03)	98.21 (0.75)
		repeated train-test split	93.39 (3.60)	0.83 (0.09)	96.77 (4.21)
Digital fingerprinting removed	A	repeated 10-fold CV	77.07 (2.01)	0.37 (0.10)	67.30 (4.55)
		repeated train-test split	75.56 (6.84)	0.39 (0.13)	69.67 (8.27)
	B	repeated 10-fold CV	74.31 (3.73)	0.29 (0.11)	67.00 (9.77)
		repeated train-test split	73.67 (10.04)	0.31 (0.28)	67.18 (20.39)
	C	repeated 10-fold CV	76.74 (2.00)	0.27 (0.09)	66.09 (8.66)
		repeated train-test split	75.47 (8.80)	0.28 (0.21)	71.62 (16.50)
Digital fingerprinting and accuracy paradox removed	A	repeated 10-fold CV	68.65 (4.88)	0.39 (0.10)	68.65 (4.88)
		repeated train-test split	68.06 (9.62)	0.38 (0.20)	68.06 (9.62)
	B	repeated 10-fold CV	65.45 (5.11)	0.31 (0.11)	70.54 (7.79)
		repeated train-test split	62.92 (10.80)	0.27 (0.23)	67.60 (13.60)
	C	repeated 10-fold CV	62.47 (4.20)	0.25 (0.10)	66.59 (8.50)
		repeated train-test split	61.81 (12.19)	0.25 (0.27)	66.53 (15.37)
All inflationary effects removed	A	repeated 10-fold CV	67.98 (4.40)	0.38 (0.09)	67.83 (4.35)
		repeated train-test split	65.18 (16.15)	0.33 (0.34)	65.06 (16.15)
	B	repeated 10-fold CV	62.45 (6.39)	0.26 (0.14)	70.33 (8.84)
		repeated train-test split	62.96 (13.10)	0.28 (0.30)	69.78 (19.50)
	C	repeated 10-fold CV	62.28 (4.62)	0.26 (0.10)	69.03 (7.94)
		repeated train-test split	61.59 (13.35)	0.24 (0.31)	64.91 (21.32)
No inflationary effects removed, but fewer samples	A	repeated 10-fold CV	90.35 (3.30)	0.75 (0.10)	88.36 (5.37)
		repeated train-test split	89.08 (5.94)	0.72 (0.15)	86.85 (8.52)
	B	repeated 10-fold CV	87.19 (2.95)	0.67 (0.08)	90.96 (4.20)
		repeated train-test split	83.79 (8.21)	0.56 (0.20)	85.76 (11.22)
	C	repeated 10-fold CV	89.51 (2.92)	0.71 (0.09)	92.67 (4.33)
		repeated train-test split	88.28 (4.67)	0.69 (0.12)	89.47 (11.81)

Table C.15
Full results with models A, B, and C with the "mPower Data Set". Accuracy and AUROC are percentages (% omitted for brevity). Standard deviation shown in brackets.

	Model	Evaluation method	Accuracy	MCC	AUROC
Unmodified (no inflationary effects removed)	A	repeated 10-fold CV	72.44 (0.07)	0.39 (0.00)	77.63 (0.05)
		repeated train-test split	72.14 (0.31)	0.39 (0.01)	77.25 (0.33)
	B	repeated 10-fold CV	72.83 (0.07)	0.40 (0.00)	78.36 (0.05)
		repeated train-test split	72.79 (0.32)	0.40 (0.01)	78.23 (0.35)
	C	repeated 10-fold CV	74.88 (0.04)	0.45 (0.00)	80.82 (0.05)
		repeated train-test split	74.85 (0.15)	0.45 (0.00)	80.65 (0.22)
Digital fingerprinting removed	A	repeated 10-fold CV	65.78 (0.47)	0.26 (0.01)	68.35 (0.49)
		repeated train-test split	67.17 (1.85)	0.27 (0.02)	69.44 (1.47)
	B	repeated 10-fold CV	69.50 (0.56)	0.34 (0.01)	74.17 (0.68)
		repeated train-test split	70.72 (2.69)	0.35 (0.04)	74.96 (2.55)
	C	repeated 10-fold CV	69.25 (0.45)	0.33 (0.01)	73.32 (0.56)
		repeated train-test split	70.65 (2.45)	0.35 (0.03)	74.08 (2.12)
Digital fingerprinting and accuracy paradox removed	A	repeated 10-fold CV	63.22 (0.34)	0.27 (0.01)	68.01 (0.53)
		repeated train-test split	63.50 (1.20)	0.27 (0.02)	68.30 (1.30)
	B	repeated 10-fold CV	66.64 (0.49)	0.33 (0.01)	73.33 (0.68)
		repeated train-test split	66.34 (1.15)	0.33 (0.02)	72.90 (1.58)
	C	repeated 10-fold CV	66.71 (0.39)	0.34 (0.01)	73.07 (0.47)
		repeated train-test split	66.74 (1.22)	0.34 (0.02)	73.14 (1.28)
All inflationary effects removed	A	repeated 10-fold CV	53.37 (0.49)	0.07 (0.01)	54.75 (0.67)
		repeated train-test split	52.76 (1.45)	0.06 (0.03)	54.09 (1.67)
	B	repeated 10-fold CV	54.55 (1.03)	0.09 (0.02)	57.26 (1.14)
		repeated train-test split	54.02 (2.18)	0.08 (0.04)	56.44 (2.27)
	C	repeated 10-fold CV	54.73 (0.64)	0.10 (0.01)	56.86 (0.86)
		repeated train-test split	54.32 (1.49)	0.09 (0.03)	55.98 (1.67)
No inflationary effects removed, but fewer samples	A	repeated 10-fold CV	70.91 (0.24)	0.36 (0.00)	75.28 (0.23)
		repeated train-test split	70.97 (0.50)	0.36 (0.01)	75.14 (0.48)
	B	repeated 10-fold CV	72.10 (0.17)	0.39 (0.01)	77.43 (0.28)
		repeated train-test split	72.07 (0.97)	0.39 (0.02)	77.43 (0.77)
	C	repeated 10-fold CV	73.80 (0.26)	0.43 (0.01)	79.32 (0.16)
		repeated train-test split	73.64 (0.54)	0.42 (0.01)	79.13 (0.57)

C.3. Additional experiments on generalization performance

The generalization performance was also investigated in the presence of inflationary effects, that is, once the generalization test set has been removed, the remaining data set is not modified before being used for model training. The results shown in Table C.16 indicate a noticeable reduction in accuracy as a result of not removing the inflationary effects.

Table C.16
Evaluating the generalization performance in the presence and absence of inflationary effects. Accuracy and AUROC are percentages (% omitted for brevity). Standard deviations shown in brackets.

Changes to the training set	Model	Accuracy	MCC	AUROC
Inflationary effects present [$n = 59,358$]	A	42.60 (5.32)	0.12 (0.08)	66.54 (8.36)
	B	45.80 (6.09)	0.14 (0.07)	68.92 (7.45)
	C	45.40 (4.90)	0.14 (0.07)	69.58 (6.83)
Inflationary effects removed [$n = 18,495$]	A	52.30 (4.30)	0.11 (0.10)	64.37 (11.39)
	B	53.80 (7.98)	0.14 (0.10)	67.33 (7.19)
	C	58.80 (5.29)	0.18 (0.07)	69.53 (9.11)
Fewer samples, inflationary effects present [$n = 18,495$]	A	42.40 (3.88)	0.17 (0.05)	72.46 (6.52)
	B	43.20 (3.79)	0.16 (0.05)	71.94 (7.34)
	C	42.70 (4.17)	0.14 (0.08)	71.42 (8.44)

Appendix D. Pseudo-code

Below outlines pseudo-code for the "virtual twins" method described in Section 3.2. Only the `main` method and the methods used in `main` are defined. The behavior of the remaining helper methods are written in plain English.

```
def main(participants):

  # get the "twins" of participants
  paired_participants = get_paired_participants(participants)

  # expand the "twinned"participants into samples
  paired_samples = participant_pairs_to_sample_pairs(paired_participants)

  # use the paired_samples for training, evaluation, testing. Each pair in paired_samples will
  # not always have the same number of samples. Splitting the data must be done with this in mind.
  evaluate(paired_samples)


def get_paired_participants(participants):
  """Finds a set of pairs of participants. """

  paired_participants = []
  participants = shuffle(participants)
  pd_participants, control_participants = split_by_class(participants)

  for p in pd_participants:
    candidate_matches = find_matches(p, control_participants)
    if len(candidate_matches) > 0:  # is not empty
      match = choose_from_candidates(p, candidate_matches)
      paired_participants.append((p, match))
      control_participants.drop(match)  # so that control cannot be matched again
  return paired_participants


def participant_pairs_to_sample_pairs(paired_participants):
  """Expands the pair of participants into pair of samples from those participants. """

  paired_samples = []
  for pd_p, control_p in paired_participants:
    pd_samples, control_samples = get_samples(pd_p, control_p)
    paired_samples.append((pd_samples, control_samples))

  return paired_samples


def shuffle(participants):
  """Randomly shuffles the participants"""
  return shuffled_participants


def split_by_class(participants):
  """From all the participants, split them into the PD class and the control class"""
  return pd_participants, control_participants




def find_matches(participant, available_participants, factors):
  """Returns a list of possible matches for a `participant` from `available_participants`, given
  a list of `factors` to consider.

  A possible match is only considered if `check_if_all_factors_are_similar` returns True for the
  participant and candidate."""
  return candidate_matches


def choose_from_candidates(participant, candidate_matches):
  """Randomly picks a match from `candidate_matches`, with the addition that the choosing is
  weighted by the similarity of the number of samples between participant and candidates.

  The more similar the number of submissions between participant and candidates, the higher the
  weight. The purpose for this is to get the most out of the data set."""
  return candidate


def get_samples(pd_participant, control_participant):
  """Gets random samples from each participant.

  The number of samples retrieved is `n = min(len(pd_samples), len(control_samples))`, where
  `pd_samples` are the samples belonging to `pd_participant`, and `control_samples` are the
  samples belonging to `control_participant`."""
  return n_pd_samples, n_control_samples


def is_similar(factor_pd, factor_control, tolerance):
  """Returns True if `factor_pd` and `factor_control` are similar, False if not.

  Similar means that, if the factors are of type `float`, tolerance is used. If factors are of
  type `str`, strict equality is used."""
  return if_factors_are_similar


def check_if_all_factors_are_similar(candidate, participant, factors):
  """Returns True if all the factors for a candidate and a participant are similar, False if not.

  Iterates through the list of factors, applying the `is_similar` method to determine if all
  factors are similar."""
  return if_all_factors_are_similar
```

References

[1] Zhang Z, Chen P, McGough M, Xing F, Wang C, Bui M, et al. Pathologist-level interpretable whole-slide cancer diagnosis with deep learning. Nat Mach Intell 2019;1(5):236–45. http://dx.doi.org/10.1038/s42256-019-0052-1, URL http://www.nature.com/articles/s42256-019-0052-1.

[2] Cao Y, Das P, Chenthamarakshan V, Chen P-Y, Melnyk I, Shen Y. Fold2Seq: A joint sequence(1D)-fold(3D) embedding-based generative model for protein design. In: Proceedings of the 38th international conference on machine learning. PMLR; 2021, p. 1261–71, URL https://proceedings.mlr.press/v139/cao21a.html. ISSN: 2640-3498.

[3] Gupta D, Julka A, Jain S, Aggarwal T, Khanna A, Arunkumar N, et al. Optimized cuttlefish algorithm for diagnosis of Parkinson's disease. Cogn Syst Res 2018;52:36–48. http://dx.doi.org/10.1016/j.cogsys.2018.06.006, URL http://www.sciencedirect.com/science/article/pii/S1389041718301876.

[4] Saravanan S, Ramkumar K, Adalarasu K, Sivanandam V, Kumar SR, Stalin S, et al. A systematic review of artificial intelligence (AI) based approaches for the diagnosis of Parkinson's disease. Arch Comput Methods Eng 2022. http://dx.doi.org/10.1007/s11831-022-09710-1.

[5] Pringsheim T, Jette N, Frolkis A, Steeves TDL. The prevalence of Parkinson's disease: A systematic review and meta-analysis. Mov Disorders 2014;29(13):1583–90. http://dx.doi.org/10.1002/mds.25945, URL http://onlinelibrary.wiley.com/doi/abs/10.1002/mds.25945.

[6] Jankovic J. Parkinson's disease: Clinical features and diagnosis. J Neurol, Neurosurg Psychiatry 2008;79(4):368–76. http://dx.doi.org/10.1136/jnnp.2007.131045, URL http://jnnp.bmj.com/content/79/4/368.

[7] Kalia LV, Lang AE. Parkinson's disease. Lancet 2015;386(9996):896–912. http://dx.doi.org/10.1016/S0140-6736(14)61393-3, URL http://www.sciencedirect.com/science/article/pii/S0140673614613933.

[8] Nutt JG, Wooten GF. Diagnosis and initial management of Parkinson's disease. N Engl J Med 2005;353(10):1021–7. http://dx.doi.org/10.1056/NEJMcp043908.

[9] Savitt JM, Dawson VL, Dawson TM. Diagnosis and treatment of Parkinson disease: Molecules to medicine. J Clin Invest 2006;116(7):1744–54. http://dx.doi.org/10.1172/JCI29178, URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1483178/.

[10] Goetz CG, Tilley BC, Shaftman SR, Stebbins GT, Fahn S, Martinez-Martin P, et al. Movement disorder society-sponsored revision of the unified Parkinson's disease rating scale (MDS-UPDRS): Scale presentation and clinimetric testing results. Mov Disorders 2008;23(15):2129–70. http://dx.doi.org/10.1002/mds.22340, URL https://movementdisorders.onlinelibrary.wiley.com/doi/full/10.1002/mds.22340.

[11] Gonera EG, Hof MV, Berger HJC, van Weel C, Horstink MWIM. Symptoms and duration of the prodromal phase in Parkinson's disease. Mov Disorders 1997;12(6):871–6. http://dx.doi.org/10.1002/mds.870120607, URL https://movementdisorders.onlinelibrary.wiley.com/doi/abs/10.1002/mds.870120607.

[12] Hawkes CH, Tredici KD, Braak H. A timeline for Parkinson's disease. Parkinsonism Rel Disorders 2010;16(2):79–84. http://dx.doi.org/10.1016/j.parkreldis.2009.08.007, URL https://www.prd-journal.com/article/S1353-8020(09)00217-X/fulltext.

[13] Hess CW, Okun MS. Diagnosing Parkinson disease. CONTINUUM: Lifelong Learn Neurol 2016;22(4):1047. http://dx.doi.org/10.1212/CON.0000000000000345, URL https://journals.lww.com/continuum/Abstract/2016/08000/Diagnosing_Parkinson_Disease.6.aspx.

[14] Media PA. Quarter of Parkinson's sufferers were wrongly diagnosed, says charity. 2019, URL http://www.theguardian.com/society/2019/dec/30/quarter-of-parkinsons-sufferers-were-wrongly-diagnosed-says-charity. Library Catalog: www.theguardian.com Section: Society,

[15] Tracy JM, Özkanca Y, Atkins DC, Hosseini Ghomi R. Investigating voice as a biomarker: Deep phenotyping methods for early detection of Parkinson's disease. J Biomed Inform 2020;104:103362. http://dx.doi.org/10.1016/j.jbi.2019.103362, URL https://www.sciencedirect.com/science/article/pii/S1532046419302825.

[16] Logemann JA, Fisher HB, Boshes B, Blonsky ER. Frequency and cooccurrence of vocal tract dysfunctions in the speech of a large sample of Parkinson patients. J Speech Hear Disord 1978;43(1):47–57.

[17] Ho AK, Iansek R, Marigliani C, Bradshaw JL, Gates S. Speech impairment in a large sample of patients with Parkinson's disease. Behav Neurol 1998;11(3):131–7.

[18] Rusz J, Cmejla R, Tykalova T, Ruzickova H, Klempir J, Majerova V, et al. Imprecise vowel articulation as a potential early marker of Parkinson's disease: Effect of speaking task. J Acoust Soc Am 2013;134(3):2171–81. http://dx.doi.org/10.1121/1.4816541.

[19] Wang M, Ge W, Apthorp D, Suominen H. Robust feature engineering for Parkinson disease diagnosis: New machine learning techniques. JMIR Biomed Eng 2020;5(1):e13611. http://dx.doi.org/10.2196/13611, URL https://biomedeng.jmir.org/2020/1/e13611.

[20] Little MA, McSharry PE, Roberts SJ, Costello DA, Moroz IM. Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection. BioMed Eng OnLine 2007;6(1):23. http://dx.doi.org/10.1186/1475-925X-6-23.

[21] Little MA, McSharry PE, Hunter EJ, Spielman J, Ramig LO. Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. IEEE Trans Biomed Eng 2009;56(4):1015–22. http://dx.doi.org/10.1109/TBME.2008.2005954.

[22] Rusz J, Novotný M, Hlavnička J, Tykalová T, Růžička E. High-accuracy voice-based classification between patients with Parkinson's disease and other neurological diseases may be an easy task with inappropriate experimental design. IEEE Trans Neural Syst Rehabil Eng 2017;25(8):1319–21. http://dx.doi.org/10.1109/TNSRE.2016.2621885.

[23] Das R. A comparison of multiple classification methods for diagnosis of Parkinson disease. Expert Syst Appl 2010;37(2):1568–72. http://dx.doi.org/10.1016/j.eswa.2009.06.040, URL http://www.sciencedirect.com/science/article/pii/S0957417409006137.

[24] Li D-C, Liu C-W, Hu SC. A fuzzy-based data transformation for feature extraction to increase classification performance with small medical data sets. Artif Intell Med 2011;52(1):45–52. http://dx.doi.org/10.1016/j.artmed.2011.02.001, URL https://linkinghub.elsevier.com/retrieve/pii/S0933365711000182.

[25] Polat K. Classification of Parkinson's disease using feature weighting method on the basis of fuzzy C-means clustering. Internat J Systems Sci 2012;43(4):597–609. http://dx.doi.org/10.1080/00207721.2011.581395, URL http://www.tandfonline.com/doi/abs/10.1080/00207721.2011.581395.

[26] Zuo W-L, Wang Z-Y, Liu T, Chen H-L. Effective detection of Parkinson's disease using an adaptive fuzzy k-nearest neighbor approach. Biomed Signal Process Control 2013;8(4):364–73. http://dx.doi.org/10.1016/j.bspc.2013.02.006, URL https://linkinghub.elsevier.com/retrieve/pii/S1746809413000359.

[27] Ma C, Ouyang J, Chen H-L, Zhao X-H. An efficient diagnosis system for Parkinson's disease using kernel-based extreme learning machine with subtractive clustering features weighting approach. Comput Math Methods Med 2014;2014:1–14. http://dx.doi.org/10.1155/2014/985789, URL http://www.hindawi.com/journals/cmmm/2014/985789/.

[28] Gök M. An ensemble of k-nearest neighbours algorithm for detection of Parkinson's disease. Internat J Systems Sci 2015;46(6):1108–12. http://dx.doi.org/10.1080/00207721.2013.809613.

[29] Ozkan H. A comparison of classification methods for telediagnosis of Parkinson's disease. Entropy 2016;18(4):115. http://dx.doi.org/10.3390/e18040115, URL https://www.mdpi.com/1099-4300/18/4/115.

[30] Caliskan A, Badem H, Basturk A, Yüksel M. Diagnosis of the Parkinson disease by using deep neural network classifier. Istanbul Univ - J Electr Electron Eng 2017;17:3311–8, URL https://electricajournal.org/Content/files/sayilar/58/3311-3318.pdf.

[31] Anand A, Haque MA, Alex J, Venkatesan N. Evaluation of machine learning and deep learning algorithms combined with dimentionality reduction techniques for classification of Parkinson's disease. In: 2018 IEEE international symposium on signal processing and information technology. 2018, http://dx.doi.org/10.1109/ISSPIT.2018.8642776.

[32] Haq AU, Li JP, Memon MH, khan J, Malik A, Ahmad T, et al. Feature selection based on L1-Norm support vector machine and effective recognition system for Parkinson's disease using voice recordings. IEEE Access 2019;7:37718–34. http://dx.doi.org/10.1109/ACCESS.2019.2906350, URL https://ieeexplore.ieee.org/document/8672565.

[33] Topol EJ. High-performance medicine: The convergence of human and artificial intelligence. Nat Med 2019;25(1):44–56. http://dx.doi.org/10.1038/s41591-018-0300-7, URL http://www.nature.com/articles/s41591-018-0300-7.

[34] Neto EC, Perumal TM, Pratap A, Bot BM, Mangravite L, Omberg L. On the analysis of personalized medication response and classification of case vs control patients in mobile health studies: The mPower case study. 2017, arXiv:1706.09574 [stat], URL http://arxiv.org/abs/1706.09574.

[35] Aich S, Kim H-C, Younga K, Hui KL, Al-Absi A, Sain M. A supervised machine learning approach using different feature selection techniques on voice datasets for prediction of Parkinson's disease. In: 2019 21st International conference on advanced communication technology. 2019, http://dx.doi.org/10.23919/ICACT.2019.8701961.

[36] Valverde-Albacete FJ, Peláez-Moreno C. 100% classification accuracy considered harmful: The normalized information transfer factor explains the accuracy paradox. PLoS One 2014;9(1):e84217. http://dx.doi.org/10.1371/journal.pone.0084217, URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0084217.

[37] Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognit 1997;30(7):1145–59. http://dx.doi.org/10.1016/S0031-3203(96)00142-2, URL https://www.sciencedirect.com/science/article/pii/S0031320396001422.

[38] Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics 2020;21(1):6. http://dx.doi.org/10.1186/s12864-019-6413-7.

[39] Sokolova M, Japkowicz N, Szpakowicz S. Beyond accuracy, F-score and ROC: A family of discriminant measures for performance evaluation. In: Sattar A, Kang B-h, editors. AI 2006: Advances in artificial intelligence. Lecture notes in computer science, Berlin, Heidelberg: Springer; 2006, p. 1015–21. http://dx.doi.org/10.1007/11941439_114.

[40] Lobo JM, Jiménez-Valverde A, Real R. AUC: A misleading measure of the performance of predictive distribution models. Global Ecol Biogeogr 2008;17(2):145–51. http://dx.doi.org/10.1111/j.1466-8238.2007.00358.x, URL http://onlinelibrary.wiley.com/doi/abs/10.1111/j.1466-8238.2007.00358.x.

[41] Powers D. The problem of area under the curve. In: 2nd International conference on information science and technology. IEEE; 2012, p. 567–73.

[42] Wang Y-X, Ramanan D, Hebert M. Learning to model the tail. In: Advances in neural information processing systems, vol. 30, Curran Associates, Inc.; 2017, URL https://papers.nips.cc/paper/2017/hash/147ebe637038ca50a1265abac8dea181-Abstract.html.

[43] Cui Y, Jia M, Lin T-Y, Song Y, Belongie S. Class-balanced loss based on effective number of samples. 2019, p. 9268–77, URL https://openaccess.thecvf.com/content_CVPR_2019/html/Cui_Class-Balanced_Loss_Based_on_Effective_Number_of_Samples_CVPR_2019_paper.html.

[44] Wooten GF, Currie LJ, Bovbjerg VE, Lee JK, Patrie J. Are men at greater risk for Parkinson's disease than women? J Neurol, Neurosurg Psychiatry 2004;75(4):637–9. http://dx.doi.org/10.1136/jnnp.2003.020982, URL https://jnnp.bmj.com/content/75/4/637.

[45] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. second ed.. 2009, URL https://link.springer.com/book/10.1007/978-0-387-84858-7.

[46] Banerjee A, Chaudhury S. Statistics without tears: Populations and samples. Ind Psychiatry J 2010;19(1):60–5. http://dx.doi.org/10.4103/0972-6748.77642, URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3105563/.

[47] Bot BM, Suver C, Neto EC, Kellen M, Klein A, Bare C, et al. The mPower study, Parkinson disease mobile data collected using ResearchKit. Sci Data 2016;3(1):1–9. http://dx.doi.org/10.1038/sdata.2016.11, URL http://www.nature.com/articles/sdata201611.

[48] Dua D, Graff C. UCI machine learning repository. 2017, URL http://archive.ics.uci.edu/ml.

[49] Boersma P, Weenink D. PRAAT, a system for doing phonetics by computer. Glot Int 2001;5:341–5.

[50] Plesser HE. Reproducibility vs. replicability: A brief history of a confused terminology. Front Neuroinform 2018;11. http://dx.doi.org/10.3389/fninf.2017.00076, URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5778115/.

[51] Bishop C. Pattern recognition and machine learning. Information science and statistics, New York: Springer-Verlag; 2006, URL https://www.springer.com/gp/book/9780387310732.

[52] Krause J, Sapp B, Howard A, Zhou H, Toshev A, Duerig T, et al. The Unreasonable effectiveness of noisy data for fine-grained recognition. In: Leibe B, Matas J, Sebe N, Welling M, editors. Computer vision – ECCV 2016. Lecture notes in computer science, Cham: Springer International Publishing; 2016, p. 301–20. http://dx.doi.org/10.1007/978-3-319-46487-9_19.

[53] Hegger R, Kantz H, Schreiber T. Practical implementation of nonlinear time series methods: The TISEAN package. Chaos 1999;9(2):413–35. http://dx.doi.org/10.1063/1.166424, URL http://arxiv.org/abs/chao-dyn/9810005.

[54] McMahon CJ, Toomey JP, Kane DM. Insights on correlation dimension from dynamics mapping of three experimental nonlinear laser systems. PLoS One 2017;12(8):e0181559. http://dx.doi.org/10.1371/journal.pone.0181559, URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0181559.

[55] Gundersen OE, Kjensmo S. State of the Art: Reproducibility in artificial intelligence. Proc AAAI Conf Artif Intell 2018;32(1). http://dx.doi.org/10.1609/aaai.v32i1.11503, URL https://ojs.aaai.org/index.php/AAAI/article/view/11503.

[56] Ozbolt AS, Moro-Velazquez L, Lina I, Butala AA, Dehak N. Things to consider when automatically detecting Parkinson's disease using the phonation of sustained vowels: Analysis of methodological issues. Appl Sci 2022;12(3):991. http://dx.doi.org/10.3390/app12030991, URL https://www.mdpi.com/2076-3417/12/3/991. Number: 3 Publisher: Multidisciplinary Digital Publishing Institute.

[57] Rusz J, Tykalová T, Novotný M, Zogala D, Růžička E, Dušek P. Automated speech analysis in early untreated Parkinson's disease: Relation to gender and dopaminergic transporter imaging. Eur J Neurol 2022;29(1):81–90. http://dx.doi.org/10.1111/ene.15099, URL http://onlinelibrary.wiley.com/doi/abs/10.1111/ene.15099, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ene.15099.

[58] Rusz J, Tykalová T, Novotný M, Růžička E, Dušek P. Distinct patterns of speech disorder in early-onset and late-onset de-novo Parkinson's disease. Npj Parkinson's Disease 2021;7(1):1–8. http://dx.doi.org/10.1038/s41531-021-00243-1, URL http://www.nature.com/articles/s41531-021-00243-1, Number: 1 Publisher: Nature Publishing Group.

[59] Cai Z, Gu J, Wen C, Zhao D, Huang C, Huang H, et al. An intelligent Parkinson's disease diagnostic system based on a chaotic bacterial foraging optimization enhanced fuzzy KNN approach. Comput Math Methods Med 2018;2018:e2396952. http://dx.doi.org/10.1155/2018/2396952, URL https://www.hindawi.com/journals/cmmm/2018/2396952/.

[60] Williams S, Relton SD, Fang H, Alty J, Qahwaji R, Graham CD, et al. Supervised classification of bradykinesia in Parkinson's disease from smartphone videos. Artif Intell Med 2020;110:101966. http://dx.doi.org/10.1016/j.artmed.2020.101966, URL https://linkinghub.elsevier.com/retrieve/pii/S0933365720312318.

[61] Pereira CR, Pereira DR, Weber SAT, Hook C, de Albuquerque VHC, Papa JP. A survey on computer-assisted Parkinson's Disease diagnosis. Artif Intell Med 2019;95:48–63. http://dx.doi.org/10.1016/j.artmed.2018.08.007, URL http://www.sciencedirect.com/science/article/pii/S0933365717305663.

[62] Tăuţan A-M, Ionescu B, Santarnecchi E. Artificial intelligence in neurodegenerative diseases: A review of available tools with a focus on machine learning techniques. Artif Intell Med 2021;117:102081. http://dx.doi.org/10.1016/j.artmed.2021.102081, URL https://linkinghub.elsevier.com/retrieve/pii/S0933365721000749.

[63] Doyle-Lindrud S. Watson will see you now: A supercomputer to help clinicians make informed treatment decisions. Clin J Oncol Nursing; Pittsburgh 2015;19(1):31–2, URL http://search.proquest.com/docview/1655557089/citation/3871F6F20E5E45A0PQ/1.

[64] Ross C, Swetlitz I. IBM's Watson recommended 'unsafe and incorrect' cancer treatments. 2018, URL https://www.statnews.com/2018/07/25/ibm-watson-recommended-unsafe-incorrect-treatments/,

[65] Strickland E. IBM Watson, heal thyself: How IBM overpromised and underdelivered on AI health care. IEEE Spectr 2019;56(4):24–31. http://dx.doi.org/10.1109/MSPEC.2019.8678513.

[66] Hu Y, Jacob J, Parker GJM, Hawkes DJ, Hurst JR, Stoyanov D. The challenges of deploying artificial intelligence models in a rapidly evolving pandemic. Nat Mach Intell 2020;2(6):298–300. http://dx.doi.org/10.1038/s42256-020-0185-2, URL http://www.nature.com/articles/s42256-020-0185-2.

[67] Rusz J, Tykalova T, Novotny M, Zogala D, Sonka K, Ruzicka E, et al. Defining speech subtypes in De Novo Parkinson disease: Response to long-term Levodopa therapy. Neurology 2021;97(21):e2124–35. http://dx.doi.org/10.1212/WNL.0000000000012878, URL https://n.neurology.org/content/97/21/e2124. Publisher: Wolters Kluwer Health, Inc. on behalf of the American Academy of Neurology Section: Research Article.

[68] Plana D, Shung DL, Grimshaw AA, Saraf A, Sung JJY, Kann BH. Randomized clinical trials of machine learning interventions in health care: A systematic review. JAMA Network Open 2022;5(9):e2233946. http://dx.doi.org/10.1001/jamanetworkopen.2022.33946, URL https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2796833.

[69] Buolamwini J, Gebru T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Proceedings of the 1st conference on fairness, accountability and transparency. PMLR; 2018, p. 77–91, URL https://proceedings.mlr.press/v81/buolamwini18a.html. ISSN: 2640-3498.

[70] Simons J, Adams Bhatti S, Weller A. Machine learning and the meaning of equal treatment. In: Proceedings of the 2021 AAAI/ACM conference on AI, ethics, and society. New York, NY, USA: Association for Computing Machinery; 2021, p. 956–66.

[71] Berk R. An impact assessment of machine learning risk forecasts on parole board decisions and recidivism. J Exp Criminol 2017;13(2):193–216. http://dx.doi.org/10.1007/s11292-017-9286-2.

[72] Peralta M, Jannin P, Haegelen C, Baxter JSH. Data imputation and compression for Parkinson's disease clinical questionnaires. Artif Intell Med 2021;114:102051. http://dx.doi.org/10.1016/j.artmed.2021.102051, URL https://www.sciencedirect.com/science/article/pii/S0933365721000440.

[73] US FDA, Health Canada, UK MHRA. Good machine learning practice for medical device development: Guiding principles. 2021, URL https://www.fda.gov/medical-devices/software-medical-device-samd/good-machine-learning-practice-medical-device-development-guiding-principles,

[74] Chen L, Suominen H. An approach to the frugal use of human annotators to scale up auto-coding for text classification tasks. In: Proceedings of the 19th workshop of the Australasian language technology association. 2021, p. 12–21, URL https://alta2021.alta.asn.au/papers.

[75] Sakar CO, Kursun O. Telediagnosis of Parkinson's disease using measurements of dysphonia. J Med Syst 2010;34(4):591–9. http://dx.doi.org/10.1007/s10916-009-9272-y, URL http://link.springer.com/10.1007/s10916-009-9272-y.

[76] Ozcift A, Gulten A. Classifier ensemble construction with rotation forest to improve medical diagnosis performance of machine learning algorithms. Comput Methods Programs Biomed 2011;104(3):443–51. http://dx.doi.org/10.1016/j.cmpb.2011.03.018, URL http://www.sciencedirect.com/science/article/pii/S0169260711000836.

[77] Sakar BE, Isenkul ME, Sakar CO, Sertbas A, Gurgen F, Delil S, et al. Collection and analysis of a parkinson speech dataset with multiple types of sound recordings. IEEE J Biomed Health Inf 2013;17(4):828–34. http://dx.doi.org/10.1109/JBHI.2013.2245674.

[78] Tang B, He H. KernelADASYN: Kernel based adaptive synthetic data generation for imbalanced learning. In: 2015 IEEE congress on evolutionary computation. 2015, p. 664–71. http://dx.doi.org/10.1109/CEC.2015.7256954, ISSN: 1941-0026.

[79] Abiyev RH, Abizade S. Diagnosing Parkinson's diseases using fuzzy neural system. Comput Math Methods Med 2016;2016:1–9. http://dx.doi.org/10.1155/2016/1267919, URL http://www.hindawi.com/journals/cmmm/2016/1267919/.

[80] Gürüler H. A novel diagnosis system for Parkinson's disease using complex-valued artificial neural network with k-means clustering feature weighting method. Neural Comput Appl 2017;28(7):1657–66. http://dx.doi.org/10.1007/s00521-015-2142-2.

[81] Aich S, Younga K, Hui KL, Al-Absi AA, Sain M. A nonlinear decision tree based classification approach to predict the Parkinson's disease using different feature sets of voice data. In: 2018 20th international conference on advanced communication technology. 2018, p. 1–2. http://dx.doi.org/10.23919/ICACT.2018.8323863.

[82] Grassberger P, Procaccia I. Measuring the strangeness of strange attractors. Physica D 1983;9(1):189–208. http://dx.doi.org/10.1016/0167-2789(83)90298-1, URL http://www.sciencedirect.com/science/article/pii/0167278983902981.

[83] Theiler J. Estimating fractal dimension. J Opt Soc Amer A 1990;7(6):1055–73. http://dx.doi.org/10.1364/JOSAA.7.001055, URL https://www.osapublishing.org/josaa/abstract.cfm?uri=josaa-7-6-1055.

[84] Fraser AM, Swinney HL. Independent coordinates for strange attractors from mutual information. Phys Rev A 1986;33(2):1134–40. http://dx.doi.org/10.1103/PhysRevA.33.1134, URL https://link.aps.org/doi/10.1103/PhysRevA.33.1134.

[85] Kennel MB, Brown R, Abarbanel HDI. Determining embedding dimension for phase-space reconstruction using a geometrical construction. Phys Rev A 1992;45(6):3403–11. http://dx.doi.org/10.1103/PhysRevA.45.3403, URL https://link.aps.org/doi/10.1103/PhysRevA.45.3403.

[86] Provenzale A, Smith LA, Vio R, Murante G. Distinguishing between low-dimensional dynamics and randomness in measured time series. Physica D 1992;58(1):31–49. http://dx.doi.org/10.1016/0167-2789(92)90100-2, URL http://www.sciencedirect.com/science/article/pii/0167278992901002.