applied sciences

MDPI

Article

The Role of Data Analytics in the Assessment of Pathological Speech—A Critical Appraisal

Pedro Gómez-Vilda [1,*], Andrés Gómez-Rodellar [2], Daniel Palacios-Alonso [3], Victoria Rodellar-Biarge [1] and Agustín Álvarez-Marquina [1]

[1] NeuSpeLab, CTB, Universidad Politécnica de Madrid, Pozuelo de Alarcón, 28220 Madrid, Spain
[2] Usher Institute, Faculty of Medicine, University of Edinburgh, Edinburgh EH8 9YL, UK
[3] Escuela Técnica Superior de Ingeniería Informática, Universidad Rey Juan Carlos, Móstoles, 28933 Madrid, Spain
* Correspondence: pedro.gomezv@upm.es; Tel.: +34-649744299

Abstract: Pathological voice characterization has received increasing attention over the last 20 years. Hundreds of studies have been published showing inventive approaches with very promising findings. Nevertheless, methodological issues might hamper performance assessment trustworthiness. This study reviews some critical aspects regarding data collection and processing, machine learning-oriented methods, and grounding analytical approaches, with a view to embedding developed clinical decision support tools into the diagnosis decision-making process. A set of 26 relevant studies published since 2010 was selected through critical selection criteria and evaluated. The model-driven (MD) or data-driven (DD) character of the selected approaches is deeply examined considering novelty, originality, statistical robustness, trustworthiness, and clinical relevance. It has been found that before 2020 most of the works examined were more aligned with MD approaches, whereas over the last two years a balanced proportion of DD and MD-based studies was found. A total of 15 studies presented MD characters, whereas seven were mainly DD-oriented, and four shared both profiles. Fifteen studies showed exploratory or prospective advanced statistical analysis. Eighteen included some statistical validation to avail claims. Twenty-two reported original work, whereas the remaining four were systematic reviews of others' work. Clinical relevance and acceptability by voice specialists were found in 14 out of the 26 works commented on. Methodological issues such as detection and classification performance, training and generalization capability, explainability, preservation of semantic load, clinical acceptance, robustness, and development expenses have been identified as major issues in applying machine learning to clinical support systems. Other important aspects to be taken into consideration are trustworthiness, gender-balance issues, and statistical relevance.

Keywords: laryngeal pathology characterization; machine learning methods; model-driven vs. data-driven approaches; handcrafted vs. end-to-end ML; AI explainability

Citation: Gómez-Vilda, P.; Gómez-Rodellar, A.; Palacios-Alonso, D.; Rodellar-Biarge, V.; Álvarez-Marquina, A. The Role of Data Analytics in the Assessment of Pathological Speech—A Critical Appraisal. *Appl. Sci.* **2022**, *12*, 11095. https://doi.org/10.3390/app122111095

Academic Editor: Sten Ternström

Received: 27 September 2022
Accepted: 28 October 2022
Published: 2 November 2022

1. Introduction

Organic and neurological disorders often leave an imprint on voice, with classical symptom manifestations broadly categorized as dysphonia and dysarthria. The development of advanced signal processing and pattern recognition tools, combined with advances in statistical machine learning (ML) methods, have produced a surge of research works, resulting in thousands of publications. Notwithstanding the complexity of the approaches, an underlying problem has not been sufficiently addressed yet. This is a very old one, which goes back to the early times of a basic controversy in Artificial Intelligence (AI): that of MD versus DD systems [1]. On the one hand, MD systems rely on specific models developed on a priori knowledge of biophysical, biomechanical, and neuromotor backgrounds behind observable correlation estimations. On the other hand, DD systems are mainly based on the data available from out-of-the-system observations, independently

of the biophysical hypotheses behind them. MD systems work under the assumption that a model that encapsulates the behavior of a system can be developed [2,3], and by changing the initial conditions and the model parameters the outcome can be predicted. DD systems work the other way around—the working model is not something to be used as a vehicle to understand the data, on the belief that data speak for themselves [4]. The complex mathematical model could be left aside as far as the behavior of the system might be inferred from the distribution of the data. Traditionally, both approaches have been assumed as opposite and competing, whereas exploring their synergies may procure new opportunities [5].

A recent example of these different approaches in pathological phonation detection can be found in [6], where results from glottal features and a support vector machine (SVM) classifier are compared with those from raw speech and glottal flow on a convolutional neural network (CNN) and a long-short term memory (LSTM) classifier, on what is known as an end-to-end (EE) approach. Implicitly, the use of glottal features assumes the need of resorting to the classical source-filter model proposed by G. Fant [7], thus the approach can be considered partially as an MD one. The use of raw speech on a CNN or an LSTM would be a pure DD approach. A system based on processing a raw glottal source on a CNN or an LSTM, also proposed in the same article, might be considered a hybrid approach, much in the way as glottal source estimation by inversion methods does, where a source-filter model is implicit. Eventually, DD systems may work more efficiently than knowledge-based systems, as they do not depend on accurate and robust systemic descriptions, and are prone to be incomplete, vulnerable, and of cumbersome implementation, whereas DD models depend on robust and compact numerical algorithms. This phenomenon has come to be generalized nowadays with the advent of deep learning methods (DL), which rely on intensive computation, proposing pure EE models based on DD systems, in contrast to MD systems, where substantial efforts must be devoted to extracting and selecting semantic features "by hand", on what has become to be known as handcrafted-feature systems (HF). This panorama would be an idealized one, and when facing real clinical scenarios the struggle would be won by EE systems if it were not for the data complexity and variability, as very often, EE systems are unreliable when it comes to generalization. Concept EE-based solutions, which work rather efficiently under very controlled conditions, are prone to fail dramatically when exposed to slightly varying situations differing from the original ones under which they were conceived and trained [8].

Careful feature processing is another problem to deal with. Complex AI system design requires recording multi-trait signals from different perspectives, including high-dimensional feature spaces, which can hinder pattern descriptions if datasets are incomplete, sparse, or corrupted. Ideally, the problem should concentrate on sets of features that maximally separate data classes, still retaining classification-related semantic meaning. If estimated clinical outcomes are expected to be continuous, regression methods will be used.

As far as voice is concerned, this problem becomes more and more complicated, because the biomarkers that can be estimated from the acoustic signal are of a broad nature [9], such as those found in phonation (including respiration, common perturbation, harmonic-noise ratios, open and close quotients, biomechanical parameters, etc.), articulation (spectrograms, cepstrograms, formant-grams, vowel space area, formant centralization ratios, kinematic distributions, etc.), prosody (energy and pitch contours, zero crossings, etc.), and rhythmic (syllable/silence counts and ratios, speech rate, etc.). The statistical description of each of these sets of observations becomes the playground on which feature estimation is based, followed by identifying robust feature subsets and selecting the most relevant ones. Traditionally in MD-based ML, the semantic load of each class is taken into consideration in feature selection (understanding by semantic load, the capability of associating each feature with a specific observed phonation behavior of clinical interest). However, in some cases where features are exclusively selected on purely statistical terms, they might lose part of their semantic load, especially when applying strongly compressing feature transformations.

In general, when developing automatic ML approaches concerning Voice Quality Assessment, it is of crucial interest to take into account the following considerations:

- Voice pathology does not involve a single concept, but many. According to the American Speech and Hearing Association [10], voice disorders may be classified into two main categories: organic (produced by alteration in the respiratory, laryngeal, or oro-naso-pharyngeal mechanisms) or functional (produced by the inefficient use of phonation mechanisms).

To help conceptually visualize the difficulties involved, two examples are shown in Figures 1 and 2, corresponding to the phonation of a sustained vowel [a:] during 4 s, the first one from a normative participant, and the second one from a patient suffering from larynx carcinoma.

These two examples illustrate the complexity of defining acoustic correlates with a semantic load informing about the nature and severity of a given phonation alteration produced by a specific pathological laryngeal disease. Organic disorders of laryngeal origin may be sub-classified into structural (physical changes or damage to mechanisms involved in phonation, such as the vocal folds, or the larynx supporting structure, muscles, and cartilages) or neurogenic (resulting from malfunction of the central or peripheral nervous system pathways to larynx, resulting in tremor, spasmodic dysphonia, or vocal fold paralysis—this might also be of unilateral or bilateral character). Functional disorders might be the result of vocal fatigue, muscle dystonia (aphasia, dysphonia, hypertension), irregular vibration modality (diplophonia), or improper phonation (ventricular). Besides, according to biomechanical behavior observation, they can be classified as induced by deficient glottal cyclicity or by asymmetrical vibration. Vocal fold-related disorders may be classified accordingly to their influence on vocal fold vibration as unilateral (polyps, cysts, edema, granulomatosis, sulci, etc.) or bilateral (noduli, papillomatosis, bilateral palsy, etc.). Systemic larynx-related disorders may include chronic laryngitis, carcinomas, etc. The critical question is that this polyhedral panorama might sometimes be ignored by publications in machine-learning laryngeal pathology characterization, although they strongly condition the behavior of the biomechanical system and the acoustical signals, which is the ground on which pathology detection and classification stand. It is really difficult to classify and even talk about a multi-perspective problem when it is not visualized, not to mention interpretability and explainability, bearing in mind that many of these disorders induce similar observable correlates (many-to-many regression and classification problems [11]).

- Many studies tend to use a small set of publicly available databases, which include a few sustained vowels only. In other words, they might not be sufficiently representative of the pathologies targeted and of the population affected, overtraining on a limited dataset, hampering generalization. In this sense, the following publicly available databases have become de facto benchmarks, among others: Saarbrucken Voice Database (SVD) [12], PC-GITA [13], Hospital Universitario Príncipe de Asturias Database (PdADb) [14], and Massachusetts Eye and Ear Infirmary (MEEI) [15].
- When using third-source databases, researchers should assess the production standards under which they had been produced to ensure that they are carefully produced regarding the use of consistent equipment, adequate acoustic settings, sound pressure levels (SPL), and fundamental frequency calibration.
- Numerous studies claim diagnostic capabilities, opening a controversial debate, as most clinicians in laryngeal pathology assessment would be skeptical of binary classifiers claiming diagnosis (for instance to decide if a patient is affected by organic or functional disorders) based on acoustical correlates only. Clinical assessment practice relies on visual and image inspection tools (for instance video laryngoscopy). What clinicians consider useful instead is a methodology to assess voice functionality during and after pharmacological or rehabilitative treatment to evaluate recovery or functional improvement. Besides, ML systems are expected to also provide information regarding what role a feature plays in treatment, and its efficiency on detection and

classification performance, as well as the risk of inducing confounding factors. The term explainable medicine (EM) is the key concept in this sense [16].

- Quite a few studies on laryngeal pathology characterization by ML methods offer some knowledge of voice production and vocal pathologies when including thoughts as "data speaks for itself", implying that such a knowledge background does not matter at all. This could be a major cause for ML rejection or unacceptability by some clinical experts and would not be doing any good to the advancement of ML applicability in this field.

Having these considerations in mind, the main aim of this study is to provide a critical overview of the most common methodological issues in the study and characterization of pathological phonation and speech, regarding the MD vs. DD methodological controversy.

Studies on phonation pathology using acoustic analysis have evolved through the past three decades from symptom-feature association methods through low-level ML disease detection and grading, to advanced ML methods based on the massive use of DL, with a limited scope on disease detection. The turning point in this progressive evolution has been the introduction of CNNs, adapting them from other fields of research as a hallmark on the way to end-to-end disease detection and characterization [6]. To track this process, the specific scope of the study has been concentrated on research published after 2010 to now, monitoring the transition from low-level to advanced ML methods.

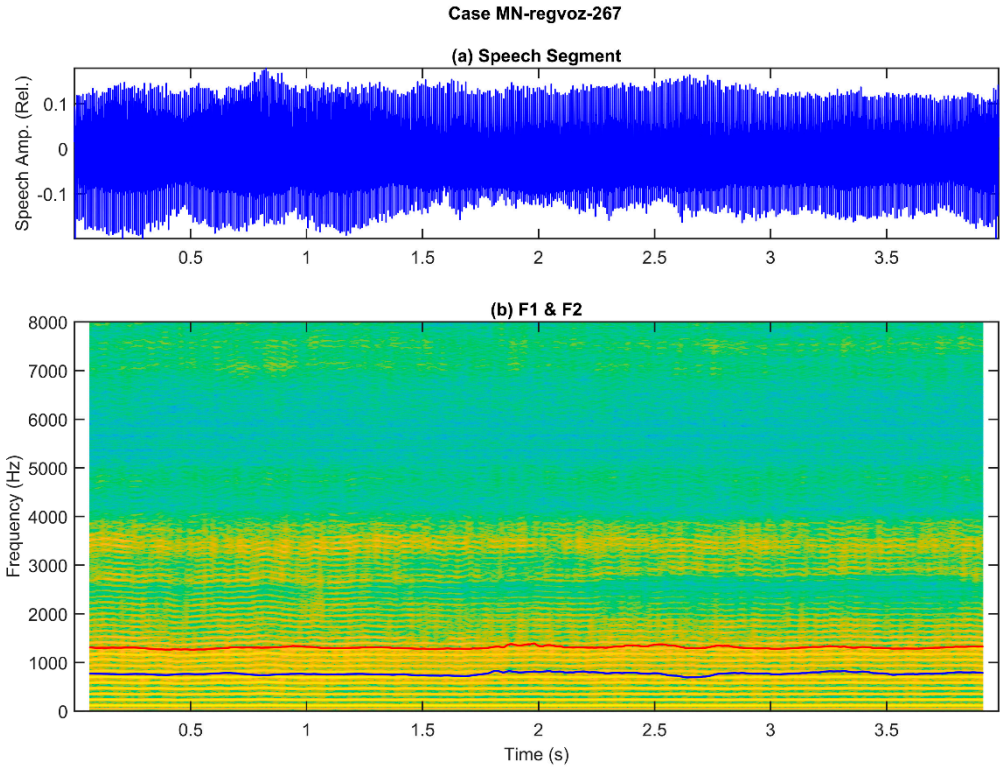

Figure 1. Example of a sustained utterance of vowel [a:] from a normative subject (male, 37 years old): (**a**) original speech signal, showing slight amplitude changes through the utterance; (**b**) spectrogram on the background and two formants in blue (first formant) and red (second formant). The harmonic spectrum (background) shows stable harmonics all over the recording, seen as horizontal yellow parallel bands, which extend from 120 Hz to almost 4000 Hz, indicating a moderate voice quality produced by a normal larynx function. The first two formants (blue and red lines) show some fluctuations, especially apparent from 1.8 s until 3.5 s, indicating a possible slight mandibular movement during phonation, as both lines maintain a relatively constant separation.

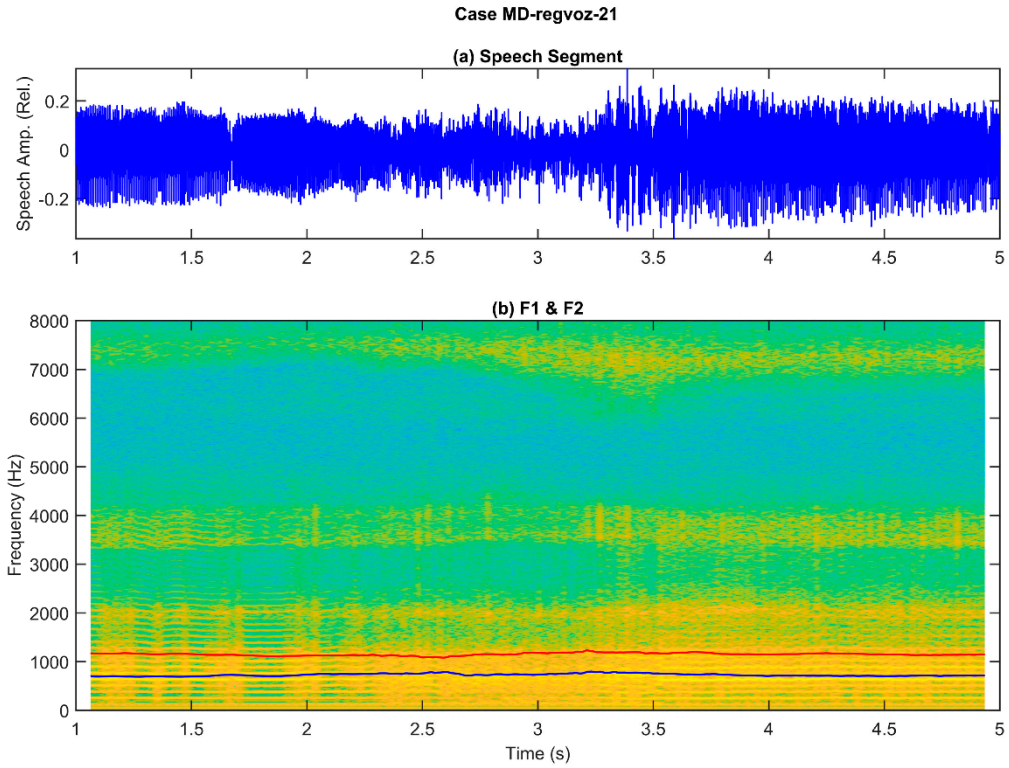

Figure 2. Example of a sustained utterance of the vowel [a:] from a patient suffering a larynx carcinoma (male, 61 years old): (**a**) original speech signal, showing strong amplitude and frequency changes through the utterance (rough and unstable phonation); (**b**) spectrogram on the background and two formants in blue (first formant) and red (second formant). The harmonic spectrum (background) shows unstable harmonics, especially from 2.3 s, seen as horizontal yellow parallel bands, showing tremor, and blurring between 2.4 and 3.5 s, which extend from 125 Hz to barely 1200 Hz, indicating a poor voice quality produced by an abnormal larynx function. The first two formants (blue and red lines) show fluctuations, especially apparent from 2.3 s to 3.5 s, indicating possible difficulties in keeping a stable oral opening during phonation, as both lines do not maintain a constant separation.

To accomplish this goal, a summary of open issues on the use of ML methods in laryngeal pathology assessment is given in Section 2. Methodological considerations regarding the application of AI to clinical decision-making, the controversy between MD or DD-oriented applications, a summary of DL methods, and the evaluation criteria used in a set of selected reviewing publications are presented. Section 3, not being exhaustive, is devoted to reviewing some of the most relevant publications, evaluated on the criteria exposed in Section 2. In Section 4, the results of the evaluations are summarized. Conclusions are presented in Section 5. References complete the document.

2. ML Methods in Laryngeal Pathology Assessment

This study is mainly concentrated on the assessment of laryngeal pathology of organic etiology on estimating the effects of asymmetry in vocal fold vibration, and closure defects, under the acoustic analysis point of view. These effects have been characterized using perturbation features in the time and frequency domain, such as jitter, shimmer, harmonic-to-noise ratio, cepstral peak prominence, long-term spectral tilt, and others similar (see [17]). Recognizing that these correlates based on linear dynamics do not describe the complex non-linear behavior of phonation, other features based on non-linear dynamics have also been proposed [18].

A critical problem in using acoustical correlates to laryngeal pathology detection and characterization is due to the non-bijective relationship between acoustical correlates and

pathological syndromes, which is known to be many-to-many indeed [19]. On one hand, a given pathology, such as a polyp, may produce similar acoustic correlates (roughness, breathiness, etc.) as a different pathology, such as a cyst. Besides, there are confounding comorbidities that may express similar acoustical correlates, for instance, an aging voice may share roughness and breathiness with other organic and neurological pathologies. On the other hand, the true added value of acoustic analysis in pathological voice and speech characterization is more focused on functional assessment (evaluating the quality of speech characteristics in terms of respiration, phonation, articulation, and prosody), and monitoring its progress in the timeline (longitudinal analysis) than to automatic diagnosis. Indeed, automatic binary detection of laryngeal pathology from acoustic analysis (if there is pathology present in phonation) is mainly confined to remote screening applications [20], because the main role of computer assessment systems would be to serve as an aid to the specialist in establishing differential diagnostic support tools, in what conceptually is computer-assisted specialist-assessed diagnosis.

2.1. Methodological Considerations

The first issue to be examined is to which extent voice disorder diagnosis itself can be granted by current AI methods. What most research publications report is binary classification on voice recordings, labeling them as pathological or not based on acoustical correlates, a result that can be barely considered a true diagnosis under clinical standards. The diagnosis of voice disorders is a rather complex process conducted by highly specialized professionals (medical doctors) applying many different analytic and exploratory methods, which include visual and instrumental inspection, such as video laryngoscopy, imagery from x-ray, fMRI, echography, electroglottography, etc., and supported by sophisticated syndrome definitions and differential diagnosis methods. Clinically speaking, what might be expected from voice disorder diagnosis should be the main label (with an associated posterior probability) selecting the potential pathology from a list of voice disorders [10], and the etiology of the pathology, conditioned to a set of features estimated not only from one but from as many acoustical recordings as possible, is assumed to be the ones better suited to describe the problem. What may be expected from laryngeal pathology assessment ML systems is a qualified and objective help estimating specific meaningful features which may be indicators of anomalous phonation, respiration, articulation, and fluency of sufficient value to address the patient to specialized services, which will be in charge of completing medical diagnosis. This has to be substantially made clear by any publication, therefore, any claim on providing automatic diagnosis from the acoustic speech signal to their ultimate implications might be taken with care, under the risk of being untrustworthy and misleading.

2.2. Clinical Approach: On Explainability

For centuries, medicine as a mainly pragmatic discipline has made a great effort to encode properly diseases, to associate specific syndromes (understood as sets of concurrent symptoms that usually form an identifiable pathological pattern), to determine etiology [21], and to create as clear as possible a guide to health disorder management, comprising diagnosis, prognosis, treatment, and rehabilitation. Therefore, any computer-assisted voice assessment system based on acoustical correlate analysis should provide adequate, intensive, and extensive explanations of the decision-making process [22], refraining from pronouncing a clinical diagnosis, for the reasons exposed above (The term *diagnosis* seems to come eventually from the Greek terms *dia* (through) and *gnosis* (knowledge), meaning that it should be practiced only through well-informed authoritative knowledge.). This requirement should be fulfilled by any AI system taking decisions implying a substantial impact on people's lives and well-being conditions, citing the EU guidelines on ethics in AI "*the right of end users not to be subject to a decision based solely on automated processing*" [23]. Explanations should include, at least, a description of the features, patterns, and categories considered in the analysis, pointing to the ones supporting the assessment in clearly ex-

pressed understandable terms for the clinical specialist, providing the descriptive statistics of the decision features, the hypotheses under test, their confidence intervals, and their associated validation [24,25]. Under this point of view, a simplistic positive/negative outcome as an answer might be insufficient concerning clinical decision-making. This requirement emphasizes the need for any serious study to be supported by a well-understood and clinically accepted biophysical model [26].

2.3. AI Approach to Clinical Decision-Taking

AI approaches to clinical practice assistance are based on the recording of multiple correlates known to be associated with specificity syndromes describing pathologies, helping in deciding on differential diagnosis. Therefore, the critical aspects to be taken into account are robust and reliable data collection, the unambiguous association among observable data and syndromes (a many-to-many relationship problem, to be solved by differencing symptom evaluation, exclusion, and reduction), significant and reliable decision-based categorical association between data and syndromes, robust management of confounding co-morbidity symptoms, and disclosing and clarifying explanation of classification decisions.

Nevertheless, a critical issue remains to be taken into consideration, namely that of clinical acceptability of any methodology linking features to classification results, providing informative tools as dashboards with feature estimations, as well as suggesting possible pathologies associated with them in probabilistic terms. In this sense, random forest tree-based methods may be more interpretable by studying feature importance, which provides some tentative insight into the likely underlying pathology through functional relationships with the features identified as most relevant ones, helping specialists to establish differential diagnosis methods relating voice disorder with etiology [27–31].

2.4. Model-Driven vs. Data-Driven Assessment Systems

The MD approach is the classical and oldest approach, maintaining that an artificial system to take decisions on recently acquired known data must be based on the knowledge extracted from a priori collected data (past historical observations, metadata, anamnesis, etc.) after a profound examination of apparent links between data classes and desired decision results deduced by logical methods such as induction, comparison, or reduction. The knowledge extracted used to be formalized in rules of the kind *if* <*condition i*> then <*decision j*> else <*decision k*>. This was in keeping with the basis of medical practice, and as such, it was massively applied in the design of medical expert systems [32]. Implementations of this paradigm using Bayesian Inference, and Fuzzy Logic seemed to work quite well, but have a complex and costly design and testing practice, consuming lots of tweaking effort, and are not free from problems related to the gap between clinician semantics and engineering practice. The term "knowledge engineering" was created to define how to bridge that crucial semantic gap [33], and gave way to the first rule-based medical expert systems, MYCIN.

The DD approach, on the other hand, is based on statistical methods to establish links between observable data and decision categorical output data, and it was very much fueled by the advent of digital computers, as it allowed the easy computation of correlation maps, conditional probabilities, likelihood associations, and complex numerical algorithms, establishing clustering relationships between observed data and desired categories or outcomes independently of rules. This approach capitalized on the seminal work of F. Rosenblatt [34] and gave rise to the term Artificial Neural Networks (ANNs), a sort of DD system, as there was no need to know the rules governing the associations, but rather only the topological links and numerical structures populating them. These systems brought a lot of new studies, reporting spectacular results in classification and decision systems, eventually leading to an asymptotic concept known as EE systems, claiming the banishing of MD systems as old-fashioned and cumbersome. This controversial debate is not new, as it can be traced to three decades ago at least [35].

Essentially, MD stand on prior formal knowledge about the problems involved in pathological phonation (physiological, biomechanical, neuromotor, and acoustical) as far as airway passages, muscles, cartilages, bones, and other related tissues are involved in voice production [36]. The Handcrafter Feature (HF) process refers mainly to the extraction of features to be used in MD systems, accordingly to their semantic load (their a priori relationship to voice disorders). In turn, some DD systems may be driven by distributed modeling as well, and therefore they could be considered hybrid approaches. DD systems may also stand on a previous HF stage, giving rise to hybrid approaches, for instance, when using glottal features (open and close quotients) to train ANNs, SVMs, or other DD systems. In their turn, it is assumed that EE systems would not require a previous HF stage. From this point of view, EE would be the ideal DD system. The relationship among these concepts is better clarified in Figure 1.

Handcrafted Feature Approach: The first stage of any AI-based detection or classification system is devoted to extracting the desired features from raw input data and well-known feature models, which will be the input to DD or MD approaches to generate the system results such as classification and performance scores. A typical example is that of mel-frequency cepstral coefficients (MFCCs) [37]. Their feature estimation model is based on windowing the acoustic signal, Fourier transform, logarithmic power spectrum, band-filter bank vectorization, and cosine transform. The types of features available for laryngeal pathology assessment are wide-ranging, basically of linear or non-linear type, and in the time, frequency, or time-frequency domains (as wavelet descriptions, for instance) [17]. The decision on which features are to be used by a given application is very much based on the previous experience and knowledge of the designer (that is where the term handcrafted features comes from), and is oriented to produce the best performance scores, depending on the loss function and the outcome for the application explored, for example, ROC curves for binary classification, confusion matrices for multi-class classification, and mean absolute error or mean squared error for regression applications. Statistical robustness and relevance are also important quality indices. Feature framing may be defined as feature estimation, followed up by feature selection to build a parsimonious information-rich dataset, using as few features as possible with maximal predictive ability. An important criterion to decide on feature selection is the semantic load of each feature considered within a joint statistical association of the selected feature subset toward maximizing the predictive power [38].

Model-Driven Approach: Model-driven approaches use formal model rules and low-middle computing-complexity algorithms based on a priori formal model rules or feature–target relationships, demanding a high parameter-tuning workload, assuming a priori knowledge about the process of feature generation. Generative methods such as Gaussian mixture models (GMMs), hidden Markov models (HMMs), or random forests (RFs) are typical classification methods used in this context. The tuning of these systems requires a certain amount of extra effort on hyper-parameter tweaking, and further back-annotation on feature estimation and classifier adjustment.

Data-Driven Approach: The main difference concerning model-driven approaches relies on the fact that the associated algorithmic models treat the data generation mechanism as unknown, not making any assumptions on the stochastic model behind feature production [39]. Discriminative methods based on support vector machines (SVMs) or artificial neural networks (ANNs) are classical classification algorithmic methods used in this context. Specifically, the success of these methods depends on the level of complexity of the data processing architectures, which may be extremely high, as in the case of deep neural networks (DNNs), composed of many hidden layers using non-linear mapping functions which may create complex discrimination patterns in the projected hidden manifolds [40]. The foundations of DNNs are similar to ANNs, with an explosion of computational requirements, which makes them training-expensive, both in floating-point operations and in the amount of data for training, depending on robust input feature databases, associated with the desired outputs (targets or labels) in a supervised process. CNNs are a subset of DNNs that use bioinspired receptive fields to extract supposedly meaningful semantic

atoms (elementary features) or micro-patterns of relevance in the hierarchical description of data structures. In essence, they are quite efficient systems if properly trained on adequate and well-populated databases. Nevertheless, these systems also require a certain amount of effort-demanding hyper-parameter tweaking.

End-to-End Approach: These systems override the need for a preliminary feature estimation stage, as input raw data are fed directly to a very complex DNN with no other feature-extraction intermediate stage, as well as the desired outputs to be associated with. The main assumption (formally unproven, although experimentally accepted) is that data speak by themselves, in a full and complex associative process grabbed during the training phase embedded in the system topology and weights, in a pure DD behavior. Similarly to DD systems, the complexity of system adjustment moves from the feature space to hyper-parameter space settings. This approach is seen by many researchers as the quintessence of current machine pattern recognition.

2.5. Collaborative Aspects of ML for Laryngeal Pathology Assessment

The review of the basic principles implied in ML will help in explaining and clarifying the most controversial issues in its application to voice laryngeal pathology detection and classification. In this sense, it is considered that DD approaches help design systems that can predict what will happen next based on what they have seen so far. Many researchers see them as "the part of AI that works". Their basic inspiration is that the data-driven way focuses on building a system that can identify what should be the right answer based on having to see a large number of examples of question/answer pairs and "training" them to get to the desired answer. The strength of these systems seems to stand on that they do not depend on human participation to accurately describe question–answer relationships by sets of explicit rules, as that is a difficult and cumbersome task. It is assumed that with DD the system learns "on its own dynamics". The more varied the training data, the better the AI system might behave. The main weakness shown by these approaches is the need for appropriate and large datasets that, crucially, must be also correctly labeled, and this is a non-negligible task, especially in the field of pathological voice assessment applications, in certain pathologies where large databases are scarce and small-sized, and prone to be confounded with other pathologies showing similar acoustic correlates. It is in this respect where these approaches are prone to fail when confronted with "live fire" in clinical scenarios, and are the ground for more-than-justified criticism [41]. Generalization capabilities are a great concern for this kind of approach [8].

MD approaches, on the other hand, present the benefits derived from attempting to capture knowledge and generate decisions through explicit representations and rules. In other words, MD represents the attempt to capture our understanding of how voice production works through accurate systemic and functional descriptions. However, there are so many different rules and exceptions that capturing the whole processes involved might be really difficult and cumbersome. Nevertheless, many voice pathologies are highly codified, explicitly defined, and can be robustly and safely captured by a model. Therefore, a strong belief exists that if a model can be carefully derived it might provide the most efficient path from question to answer. To put it otherwise, is not there any result coming out from precedent decades of research in logics, planning, semantics, or expert systems that could help us build better autonomous AI systems [42]?

There may be situations that are a good fit for the explicit application of DD or MD approaches where an additional gain may be derived from combining the two. Recently, a movement has appeared in favor of combining DD systems with MD systems to improve, complete, and make AI applications more explainable, ethical, and respectful of user needs and rights [43]. As an example, consider using a DD approach to provide abstract vector sets from raw speech records (generic layer) and combining them with an MD approach based on efficient random forests of decision tree classifiers using differential diagnosis from clinical rules (domain layer). This would be exactly where hybrid DD-MD AI would get the best of both approaches, associating powerful abstract datasets to ontological

domains, inference, and planning. Besides, from within AI research, agent-based software development can help in explaining a system as a whole and the interactions between the various subsystems involved. Agent-based software development deals with exactly these issues. For instance, it would be possible to formalize clinical knowledge on ontological domains of speech pathologies (terms, syndromes, treatments, rules, interactions, and relationships), creating a relational map based on Bayesian inference trees (like in forensic sciences [44]), and building top-down → bottom-up AI models for diagnosis-oriented platforms, supporting clinical acceptability.

2.6. MD vs. DD Comparison Criteria

To help in contrasting and adding value to the controversial and collaborative aspects of MD- vs. DD-based ML applications to clinical decision-making, the following is a list of the most relevant criteria to compare both approaches:

- **Detection and classification performance**: The ability to detect pathological from non-pathological behavior in phonation, given in universally accepted scores, such as sensitivity, specificity, accuracy, area under the ROC curve, and other similar quality standards.
- **Training capability**: The ability to create internal representations from benchmark databases universally accepted as valid standards in reasonable and well-defined termination conditions, granting scalability.
- **Generalization capability**: The possibility of maintaining the classification performance from training when new unseen datasets are presented to the classifier.
- **Explainability**: The capability of associating performance quality scores with pathology syndromes, offering clinically interpretable indications on which syndromes can be associated with specific classification scores.
- **Semantic load**: The capability of associating performance quality scores to a priori established specific study hypotheses, allowing their acceptance or rejection in formal terms, extending to new hypotheses from observed explained results.
- **Clinical acceptance**: The ability to produce performance quality scores that might be adapted to the dashboards used in voice disorder clinical scenarios.
- **Computational costs**: The possibility of supporting the training, testing, and validation phases on reasonable standard computing platforms and servers, limiting computation power and memory requirements to reasonable terms.
- **Portability**: The capability of deploying computing solutions on limited portable devices, for instance, using tiny ML techniques.
- **Robustness**: The ability to achieve reasonable performance facing scarce, noisy, or low-quality audio signals.
- **Development effort**: The efforts demanded system deployment, testing, and evaluation to satisfactory performance standards.

These criteria will help in building a listing of advantages and disadvantages favoring one approach or the other, as in Table 1. The observations presented in the table respond to general features of behavior, not to specific implementations, which may be subject to other special design criteria.

Table 1. Comparison of DD vs. MD approaches. Observations: [1] If contaminated by training data (see [8]). [2] Assuming that results meet specific clinical criteria. [3] When using transfer learning (TL) with little adaptation needs.

Criteria/Orientation	Data-Driven	Model-Driven
Performance	High	High
Training	High	Moderate
Generalization	Low [1]	Moderate
Semantic power	Low	High
Explainability	Low	High
Clinical acceptance	Low	Moderate [2]
Computational costs	Large	Low
Portability	Poor	Moderate
Robustness	High	Poor
Development Effort	Small [3]	Moderate

3. Review of Selected Works

This section is devoted to presenting and discussing some of the most characteristic works hitherto published in voice laryngeal pathology assessment using either MD or DD approaches, or both. Relevant publications in this field appearing before 2010 have been omitted for the sake of brevity, as they can be found in the bibliography of any of the reviews listed in Table 2. The publications selected and discussed here have been examined under the following classification criteria, which will be further extended in the next subsection:

- Related to voice laryngeal pathology characterization using acoustical signals.
- Having been published later than 2010.
- Describing explicitly the classification or regression methodology used.
- Using standard databases or Supplementary Materials to allow reproducibility.
- Granting results and claims supported by reliable performance metrics.
- Contributing novel and original work of sufficient merit recognized by cross-citation.

3.1. Study Assessment Premises

The criteria taken into account in the selected studies on voice laryngeal pathology characterization listed in Table 2 aim to answer what a study in laryngeal pathology assessment is expected to provide, whether based on MD or DD approaches regarding the following premises:

Trustworthiness: A scientific study is required to provide a clear description of experimental settings for third parties to be able to check results by manifesting the study hypotheses and describing the methodology in easily understandable means, either as formal, mathematical, or open code descriptions, or providing them as Supplementary Data or in code repositories. The study should make use of standard databases well-accepted by the research community, at least for reference purposes, even if proprietary databases were used.

Gender issues: Given the nature of voice, the data subsets used in the studies should be well-balanced and populated, both in gender and age, matching paired distributions between normative and pathological subsets, whereas experiments should consider gender-separated studies or gender-sensitive feature-normalization procedures. This strong requirement is based on clear reasons: human voice acoustic characteristics are strongly dependent on sexual hormone activity throughout the lifetime, especially in the case of the female voice (childhood, puberty, menstrual period, pregnancy, menopause [45,46]), showing strong differences in phonation and articulation concerning males, who on their counterpart also experience strong changes during lifetime [47] (from childhood to third age with a remarkable transition during puberty [48]). An example of pathology detection degradation when gender separation was not taken into account is reported in [49]. If studies do not consider separate samples and populations by gender and age, there is

the risk that normative male features might be taken as non-normative or pathological ones when classified under the female normative feature distributions, and vice-versa, introducing important bias affecting classification score trustworthiness, and hampering acceptability by clinicians and users: "The design of the majority of algorithms ignore the sex and gender dimension and its contribution to health and disease differences among individuals. Failure in accounting for these differences will generate sub-optimal results and produce mistakes as well as discriminatory outcomes" [50]. Taking this matter into account, governments and regulatory agencies are producing important legal requirements to adopt the best methodologies and practices for AI and ML in medical products and services to be respectful of gender bias issues (see [51] for a comprehensive review).

Statistical relevance: Study hypotheses should be stated in statistical terms and methods, declaring confidence intervals, providing supporting statistical tests, and *p*-values when possible and if meaningful. For instance, in binary classification cases, results could be presented in terms of ROC curves handling small-stride threshold fixing if possible (on smooth many-point curves). Multiple-class separation problems could use performance metrics such as Cohen's Kappa, or multi-class F-measures [52–54].

Novelty: Original studies are expected to propose new methods, features, databases, or other related issues conveying real advances in the field. Simply extracting classical features from well-known third-party databases, on third-source ML platforms capitalizing on TL, adding some slight modification to claim a small outperformance on others' work, failing to prove trustworthiness and statistical relevance avail, may not be found acceptable for consideration, as they are quite subjective and of little meaning to the research community. This may seem a very controversial issue, but it is part of this critical appraisal, as in scientific work, recognition comes from relevance. In this same sense, a deep reflection to be brought to consideration here is the use of apparently new terminology in this field, the fruit of a shallow review of the state-of-the-art, ignoring many times, and in many cases, prior well-established lore. An example of this kind is the case with "Time Distributed" NNs, which is a very old concept, dating back to the pioneering work of Waibel and Hinton [55]. A similar discussion is that of Extreme Learning Machines (ELMs), which are new terms for old concepts working on the pioneering use of Moore-Penrose Pseudoinverse, known for their single-lap training process, and which have been proposed to substitute some of the training steps of intermediate convolutional layers [56,57].

Clinical explainability: Research on laryngeal pathology assessment should be intended to express results in terms of semantic correlates related to vocal fold vibration anomalies, such as closure defects and asymmetry, to associate system–behavior complex features to acoustic perceptual phenomena, give scores as probabilities or odds better than binary true/false outcomes, and to explain comorbidity factors. Emphasis should be placed on the relationship of specific features with detection and classification outcomes.

Self-criticism: A very important aspect of any research work is to be open to criticism and discussion, these being two fundamental pillars of the scientific method. This process should be started by the authors themselves in their manuscript, where reflections on the limitations of their study, and possible improvements should be addressed. For instance, if not all the above-mentioned criteria and requirements are met, it is highly recommended that they should mention the reasons, and why and how they might affect the trustworthiness and relevance of the results in a separate section under the title Study Limitations.

3.2. Paper Selection Criteria

Due to the large number of publications produced during the past years, only specific works describing the most relevant contributions are selected following the criteria mentioned before, as older work is already well known to the research community. The paper selection was conducted on well-known databases, such as Scopus and PubMed, using the keywords laryngeal pathology, voice pathology, phonation pathology, and speech pathology, combined with acoustic analysis, and assessment, excluding functional and

neurological etiology. Only papers published in indexed journals were taken into consideration. An impact index being estimated as the average number of citations in the years since the publication was used, selecting those papers with a minimum index value of two. The selected works were evaluated to meet the following characteristics:

- **Exploratory or prospective advanced statistics (EPAS)**. Detection and classification methodologies must rely on robust statistical foundations. Separability and clustering properties rely strongly on feature distributions, connected with syndromic descriptions because differential diagnosis as practiced by specialists depends strongly on the reciprocal association of features linked to symptoms bi-directionally and in one-to-many and many-to-one relationships. In a further effort to ensure interpretability, features should be functionally validated in the context of their relationship to meaningful symptoms. Clinical studies very often have a clear exploratory or prospective character involving small datasets, not of enough size to claim statistical significance but which might gain insight from their assessment against normative databases, either using classical hypothesis tests from descriptive parametric or non-parametric methods, or advanced comparisons based on information theory, higher order statistics [58], or non-linear dynamical analysis [59]. EPAS methods help in offering the first view on pathological etiology, disease understanding, functional evaluation, and monitoring, aiming at an accurate diagnosis, management, treatment, and rehabilitation, to be further verified and generalized on extended-size databases.
- **Statistical validation of performance (SVP)**. Many good studies decay in interest if performance results are given as simple scores, lacking statistical support as *p*-values or significance intervals. Supporting results on more specific performance data (such as ROC or DET curves, in the case of binary detection, on normalized multi-class performance measurements), may help in giving a clearer idea of the reliability of results, and their acceptability, supporting the confidence of not being produced by chance. Using pure ML data analytics without the support of appropriate statistical validation, no matter how well-performing ML methods may be, could lead to difficult and unrealistic clinical applicability. This is a common problem found in many studies analyzed in the preparation of this critical appraisal not selected for failing to meet this condition. Working on second-hand databases, on a problem modeled only from a shallow point of view, or using transfer-learning methods capitalizing on others' results obtained on a very different background by transfer-learning-based ML methods, might lead to apparently impressive results, which would not reproduce well when exposed to a real clinical scenario. The ease of second-source data access and simple re-training of off-the-shelf pre-trained ML products to generate higher scores lacking sufficient statistical support [60] might not contribute a great deal to the real significance and advancement of voice pathological study and comprehension. Research reported in [27,61] may exemplify good practice regarding EPAS and SVP as quality criteria. A different field far from voice quality analysis may offer inspiring methods illustrating how to report results from conventional and metaheuristic training [62], and on subset simulation to inform on results from unsupervised learning by a divide-and-conquer approach [63].
- **DD vs. MD character (DD/MD)**. As has been explained before, this is a very relevant methodological characteristic, as it may imply a trade-off between cumbersome work either in the feature domain or in the classification design domain. No matter how excellent DD algorithms may be, they may suffer from overfitting or poor generalization, especially when the dataset size is unbalanced, or not large enough. Although data augmentation can in part cope with the computational problem, the representation of a rather rich space of possibilities in the replication of a few samples might not reproduce the real world of pathological phonation behavior in full. Model-driven approaches, although initially cumbersome, may offer a more specific view of pathology, assuming that at least some representative cases from each class may be at hand and that a differential protocol may be established based on statistical inference. Neither

method will be perfect, and an MD-based preliminary (handcrafted) approach may produce preliminary reasonable results, which could be further refined by DD-based approaches at a later stage.

- **Clinical relevance** (CR). This is a major issue and one which should be carefully taken into account to grant the success of AI-based precision medicine [30,41]. Any methodology failing to provide a clear interpretation of results in clinical terms will barely be accepted by specialists. The success of ML methods in medical applications seems to be linked to their ability in preserving explainability [64]. In this sense, it must be taken into account that explainability may be seriously hampered by certain complexity reduction methods, strong feature selection procedures, non-specific data augmentation, or inadequate ML choice between classification vs. regression approaches. A summary of good practices in this respect may be found in [23].
- **Original research vs. state-of-the-art review (O/R)**. This study has taken into consideration both types of publications. Original contributions should help in providing relevant methodological advances or new problem envision. Those reporting only slight improvements on scores re-using overly well-known methods on data samples not considered relevant enough, or not supported by statistical evidence-building procedures have been disregarded, as these works inflate the size of the state-of-the-art at the cost of not offering much new. Review studies are also most of most relevance to help emphasize well-documented original studies.

3.3. Review Results

In this section, a critical review of different ML publications on laryngeal pathology assessment is presented. Intensive and comprehensive reviews of different approaches to pathological voice assessment may be found in [65–68]. Interesting recent original research on multimodal and poly-syndrome MD approaches can be found in [17,18,28,69–76]. Recent approaches in the domain of DD pathological voice assessment are worth being mentioned [6,31,77]. Table 2 gives the list of the selected publications and their qualification concerning the characteristics mentioned (marked as "X" when fulfilling EPAS, SVP, or CR, "MD", or "DD" for DD/MD, and "O" or "R" for O/R).

Table 2. Pathological speech and voice evaluation approaches reviewed in the present study (Selected accordingly to the criteria before mentioned: addressing laryngeal pathology characterization from acoustical signal analysis, having been published since 2010, disclosing classification or regression methodology, based on standard databases or supplementary materials to allow reproducibility, supported by reliable performance metrics, and contributing novel and original work.). EPAS: Exploratory or prospective advanced statistics; SVP: Statistical validation of performance; DD/MD: Data-driven or model-driven (the label within parenthesis indicates the strategy of the approach regarding Figure 3, as a, b, c, d, or a combination); CR: Clinical relevance; O/R: Original contribution or review.

Reference	EPAS	SVP	DD/MD	CR	O/R
Arias et al. [61]	X	X	DD (c)	X	O
Uloza et al. [27]		X	MD and DD (b,c)	X	O
Akbari and Arjmandi [70]	X	X	MD (b,c)	X	O
Mekyska et al. [17]	X	X	MD (a, b)	X	O
Orozco et al. [69]	X	X	MD (a, b)	X	O
Verikas et al. [28]	X	X	MD (b)	X	O
Hemmerling, Skalski, and Gajda [74]			MD (b)	X	O
Moro, Gómez, and Godino [71]	X	X	MD (a,b)		O
Al Nasheri et al. [20]		X	DD (a,c)		O
Travieso et al. [18]	X	X	MD (a,b)		O
Verde, De Pietro, and Saninno [72]	X	X	MD (b)	X	O
Gómez-García et al. [66]			MD (a, b)		R

Table 2. *Cont.*

Reference	EPAS	SVP	DD/MD	CR	O/R
Hegde et al. [68]			MD and DD (b,c)		R
Kadiri and Alku [78]	X	X	MD (a,b)	X	O
Barreira and Ling [75]	X	X	MD (b)		O
Chen and Chen [79]			DD (a,c)		O
Harar et al. [31]	X	X	DD (c)	X	O
Islam et al. [67]			MD and DD (a,b,c)	X	R
Mohammed et al. [80]			DD (c)		O
Narendra and Alku [6]	X		MD and DD (b,c,d)		O
Wu et al. [81]	X	X	MD (a,b)	X	O
Ding et al. [77]		X	DD (a,c)		O
Lee [65]		X	DD (a,c)		R
Pützer and Wokurek [73]	X		MD (a)	X	O
Omeroglu, Mohammed, and Oral [76]		X	MD (b,c)		O
Zhou et al. [82]	X	X	MD (a,b)	X	O

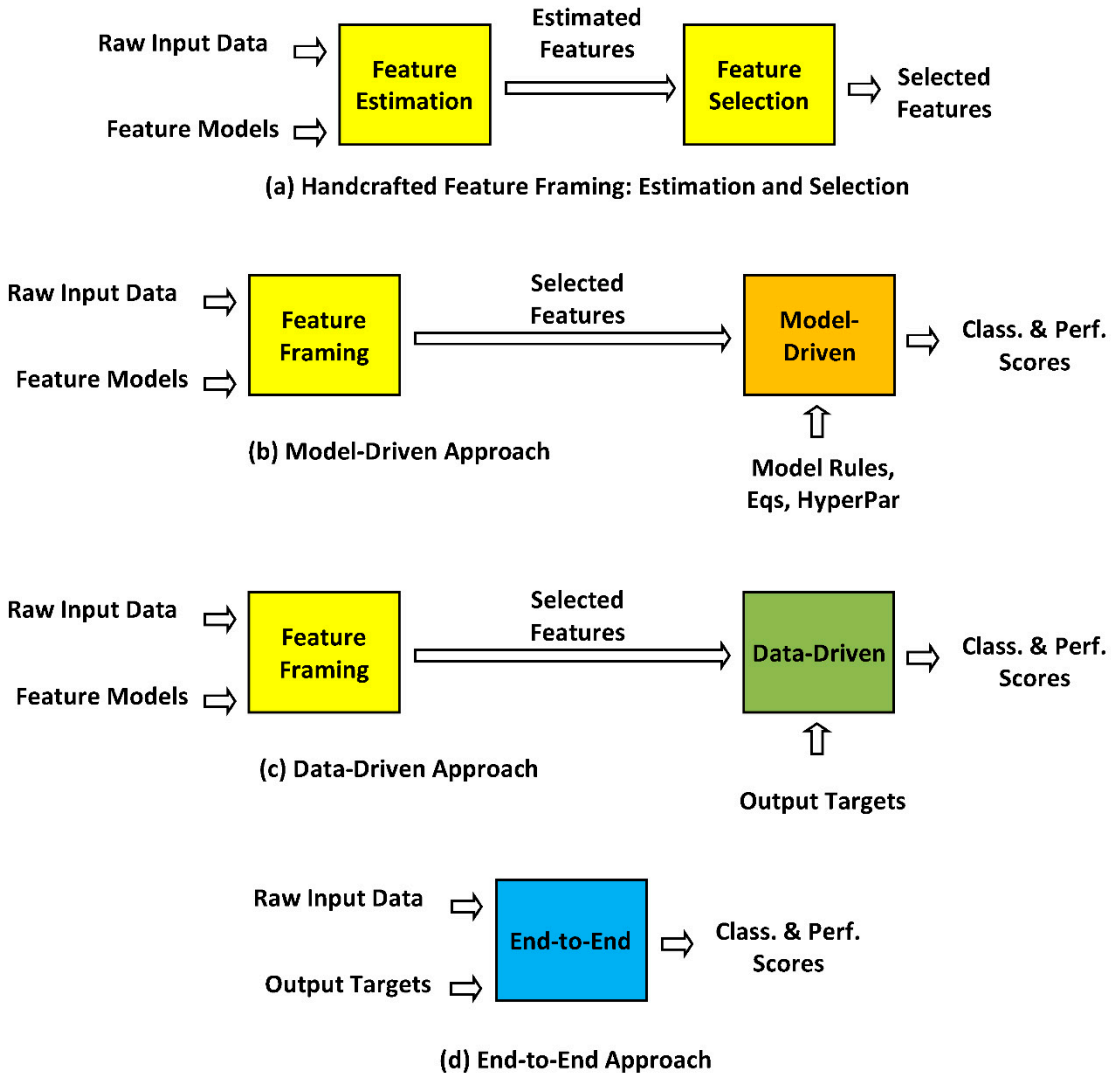

Figure 3. Different strategies for ML pathology detection and classification: (**a**) Handcrafted feature framing, comprising feature estimation followed by feature selection, based on physiological and acoustical models of voice production to estimate input features; (**b**) Model-driven approach, using features extracted by (**a**) and detection and classification model rules and mathematical structures

such as Gaussian mixture models, random forests, shallow neural networks, support vector machines, etc. to estimate classification and performance scores; (**c**) Data-driven approach, also using features modeled by (**a**) and detection and classification by deep learning methods as deep neural networks, associating expected output targets on intensive numerical computation to estimate classification and performance scores; (**d**) End-to-end approach, where stage (**a**) is omitted, using direct raw input data and desired output targets to produce classification and performance scores. Classically, MD (**b**) and DD (**c**) approaches take input features provided by a previous HF stage (**a**). End-to-end approaches do not require a previous HF stage (**a**).

4. Discussion

4.1. Analysis of Results

The works summarized in Table 2 have been listed chronologically, highlighting the fact that before 2020 most of the works showed a marked MD character, whereas since then, works associated with DD character came to a balance with those labeled as MD. In this sense, 15 would present a clear dominant MD character, whereas seven are mainly DD-oriented, and four might share both. Of course, the decision of labeling work as MD or DD is biased by subjectivity, therefore readers are challenged to do their classification. Other relevant observations are that the number of articles, including some kind of exploratory or prospective advanced statistical analysis, is 15 out of a total of 26. Similarly, 18 out of 26 publications include some form of statistical validation of performance. Interestingly, clinical relevance (CR) is present in 14 out of the 26 works studied. Finally, 22 articles reported original work, whereas the remaining four reviewed others' work using their systematic meta-analysis.

Some interesting observations might be drawn from the analysis of the works selected and commented. On the one hand, extra effort will be required to devise handcrafted pipeline feature extraction in MD approaches, as well as for feature selection, the selection of adequate statistical evaluation and classification methods, and for model and dataset re-adaptation and revision. A relevant issue is the capability of solutions for escalation and generalization. Besides, special care has to be devoted to feature selection side effects, as explainability and interpretability might be seriously affected otherwise. As a matter of fact, data high-dimensionality problems must be handled with special care, at least in different feature spaces, such as those associated with voice production, respiration capacity, phonation distortion, signal-to-noise ratios, glottal function adduction, abduction, and closure, vocal fold dysfunctional stiffness, and asymmetrical oscillation, etc. This dysfunctional behavior is usually represented by parametric and non-parametric features, such as means, standard deviations, quartiles, interquartiles, maxima, minima, and their amplitude distributions. Dysarthria-related articulation problems are usually described by spectrograms, cepstrograms, formantgrams, vowel space area estimates, formant centralization ratios, kinematic descriptions, log-normal distributions, etc. Energy and pitch contours, noise–harmonic ratios, zero-crossing estimations, syllable duration, and voiced–unvoiced ratios are used as prosody and rhythm features. Non-linear behavior features are based on fractal descriptors, phase–space attractors, and Lyapunov exponents, among others. Consequently, this rich feature environment, involving data from quite different characteristics might require special feature selection procedures to decide on inclusion, exclusion, or feature compression, and rigorous methodological classification. On the other hand, given that DD approaches show high structural regularity, which makes them especially suited for highly parallel processing at the expense of large-size data sets, and adequate DL hyper-parameter adaptation and processing methodologies' selection. Although there is a movement in favor of foreseeing DD approaches surpassing experienced clinicians in pathological voice quality detection, it seems that most published approaches rely on proprietary databases not publicly available, and the assumptions used in experimental framework design make them unsuitable for clinical application. Many such cases reviewed were based on small unbalanced sample-size databases to grant reliable performance, especially when dealing with pathologies labeled under multiple classes, and were not included in the present study. Dataset scaling presented an added problem, as in most of the cases the classification

methodology, would require model re-training from the scratch, as the solutions provided were strongly dependent on the specific classification models and databases used, raising reasonable doubts about their potential performance when tested on more general scenarios, this being the case of some EE approaches. This concern calls upon the growing need for benchmarking databases, providing larger and wider datasets covering multiple voice pathology cases to approach the problem of detection and classification meeting reliable clinical standards.

A common problem affecting MD and DD approaches studied is that of data preprocessing, comprising initial data cleaning (removal of unreadable data structures), elimination of gaps (imputation), categorical feature encoding, data normalization, and outlier removal, among other issues. Although this problem affects mainly big databases, it is seldom seen in laryngeal pathology assessment, and besides the automatic solutions that have been proposed in this sense, the issue remains one to be concerned about [83,84]. In any case, these considerations might put under question the claim that EE approaches are of immediate application and free from tweaking or handcrafting.

As a consequence of the review results and observations, the following methodological recommendations availing good practice procedures are to be taken into account helping to ensure that robust, reliable, and efficient experimental frameworks are granted for the research in the field of laryngeal pathology assessment:

- **Size and balance of the databases**. Sample size should support reasonable statistical relevance to results. Demographic information should be provided (age, gender, co-morbidities). Inclusion and exclusion criteria are to be explicitly declared. Age matching between pathological and control participants must be ensured. Equal or balanced representation of pathology sample sizes should be sought to prevent majority/minority skew and other undesired effects. Special care has to be taken if certain data-enhancing techniques, such as data augmentation, are implemented. In this sense, generative adversarial networks (GANs) might provide a way to learn deep representations without extensively annotated training data, as an alternative to other data augmentation methods [85,86].
- **Gender separation**. Special care needs to be taken in this regard, considering males and females separately as far as feature extraction and pattern matching are concerned, or using instead gender-sensitive feature normalization methods. This requirement is based on the sexual dimorphism conditioning the physiology of respiratory, phonation, and resonance systems strongly [45,46], and consequently introducing important skewness in features, which may lead to erroneous or inaccurate results. For the same reasons, when a high-performance demanding analysis is sought, as in singing voice, data have to be separated according to phonation mechanisms (as M1/M2).
- **Database recording methods**. Instrumentation, especially in microphone characteristics, room acoustics, recording levels, background noise, digitalization, downsampling, power-line filtering, and other disturbing factors, are to be carefully documented [87]. Procedures should have to be certified according to agreed criteria, e.g., Level 0 = unspecified, Level 1 = acceptable for perceptual assessment only, Level 2 = acceptable for fundamental frequency and level analysis with low noise, Level 3 = highest quality, anechoic, amenable for sensitivity analyses.
- **Methodological issues regarding feature extraction**. Feature semantic load is to be sought and preserved, and feature selection must contribute originality to the research, provided that significant improvements are shown not only from the ML perspective but also from the clinical explainability point of view.
- **Problem orientation**. Clarifying the main aim of problems treated, whether that be detection, classification, prediction, regression, longitudinal monitoring, etc. The objectives of the study must be clearly and accordingly fixed to one or several of the mentioned problems.
- **Clinical acceptance**. There must be a clear applicability orientation to clinical semantics, and efficient computer-assisted monitoring, as well as to treating confounding factors.

- **Ethical implications, and patient and data protection**. Evidence should be shown on compliance with ethical standards. The origin, protection, and protocol implementation requirements have to be explicitly declared, even when using second-source databases.
- **Information integrity**. Possible information content alteration due to too strong feature selection algorithms must be well explained. Feature selection algorithms used in data pruning must be explicitly commented on (Fisher's linear discriminant analysis, relief feature selection, embedded random forests, least absolute shrinkage, and selection operator, etc. [88–90]).
- **Linguistic issues**. The relevance of different linguistic origins of the databases should be taken into account. Although the effect of this difference could be considered negligible as far as open vowel phonation is concerned, it should not be the case when other speech tests are considered.
- **Relevant metadata**. The demographic description of both pathological and control participants, is another question to be taken into account. Inclusion and exclusion criteria have to be declared. Medication and other perturbation factors must be commented on.
- **Model selection and description**. Decisions on this important issue have to be documented, specifying the type as supervised vs. unsupervised approaches. The classification models have to be described or properly referenced, as well as the problems found in hyper-parameter settings, training, testing, validation, and evaluation.
- **Objective performance improvements**. The studies should show real significant performance improvements or knowledge advancement in the field. Orienting the research to simply outperform others' scores showing slight subjective or not statistically supported gains would not be considered sufficient for publication by many journals unless providing clear quality, innovation, and work interest.
- **Preservation of semantic load and future hypothesis casting**. The risks of using too exhaustive and tight DD and EE approaches not caring about preserving feature semantics may be hampering scientific reasoning, explainability, and the production of new hypotheses and testing paradigms. The belief that "data speak from and by themselves" is not only to be declared but also explained and proven.

4.2. Final Remarks

Some additional reflections on these methodological issues close this section. Currently, many publications are based on the use of TL from CNNs trained on visual databases on voice and speech processing. In doing so, it may be discussed if a pre-trained structure using vast amounts of visual data is the most convenient approach, even if their input consists of audio-based time-frequency spectrograms. The question is if all pathological features present in speech are adequately represented by visual receptive fields included in pre-trained CNNs, and if the large amount of data initially conceived to support image applications is not too misadjusted for audio applications, overloading training, and testing processes. It must be questioned if the basic pattern information present in the visual receptive fields at the convolutional stage, dragged down from image processing and recognition TL is the most adequate one to cover the purposes of voice and speech pathology characterization. In other words, there are unique and non-confounding separable "visual pictures" specific to each major pathology (organic, such as nodules, polyps, edema, laryngitis, carcinomas, etc., or functional, such as paresis, paralysis, neurogenic, etc.). To put it otherwise, respiration, phonation, articulation, prosody, and fluency are encoded in time-frequency spectrograms for their characteristics to be captured by visual receptive fields in CNNs trained on image databases.

Acknowledging the multi-faceted character of the different main traits of speech production (respiration, phonation, articulation, prosody, fluency), it is not unwise to think that they could be better captured by hybrid DD and MD systems, rather by one or the other isolated. This consideration could inspire cooperative agent-based software

approaches. Each macro-agent would capture common pathology patterns from each main trait and display them in semantical dashboards, helping in clinical explainability terms. For instance, auditory receptive fields might be the basic agents regarding articulation, programmed once to cover different strides within the same dimension, embedding all possible pathological patterns (FM units) within a single class in a time distribution strategy. Citing [42]: "In a more general sense, we need to think about the entire system as made up of a variety of sub-systems. Some sub-systems will be data-centric while others will depend on explicit models. While the current focus on the data-driven layer is understandable, a wider focus on the entire system is necessary". In this sense, some relevant improvements on ML systems applicable to laryngeal pathology assessment in clinical scenarios would take into account a good knowledge corpus built up during these last two decades, which has given rise to important aids to produce robust and reliable feature databases from different corpora [91]. The use of these supportive tools could ease the design of pre-processing stages. A quite relevant initiative to help in improving the application of ML and EPAS methods providing significant clinical advancements would be to create a benchmark database of speech pathologies covering organic, functional, and neurological etiologies, multiple utterance types, age and gender diversification, timely longitudinal, supported by rating scales, including multi-trait side-effect recordings (EEG, EGG, sEMG, 3DAcc, Actigraphy, etc.), and patient meta-information (posology, mental state, depression, and anxiety, co-morbidities, medication effects, aging, etc.). In dealing with small databases from specific pathologies, data augmentation could provide a practical solution for small-size databases of pathological speech [86], as well as generative adversarial networks (GAN), at the expense of an extra designing effort, and possible instabilities [92,93]. As it seems clear that DD and MD approaches present advantages and inconveniences when confronting certain applications (for instance, DD behaves quite well in image processing, segmentation, classification, etc., whereas MD systems such as ensemble methods, bagging decision trees, random forests, etc. appear to be more suitable choices for clinical explanatory diagnostic-help applications), the development of hybrid agent-based applications could benefit from the best of both approaches. Using multilayer extreme learning machines, known for their single-lap training process, could substitute some of the training steps of intermediate convolutional layers [56,57]. Presentation of results on clinically-oriented dashboards, helped by decision-making assistance on random forests, seems to be quite an acceptable approach to clinical explainability (see the work of Wu et al. [81] as an example).

It must be stressed here that the interpretability of ML results is a major issue. If asking an ML system for results, one may get scores and accuracies, but besides that, a further explanation to understand the ultimate reasons which led the ML system to produce the results shown is needed. This requirement, which is known as explainability, is one of the hot issues in AI nowadays. Explainability is related to semantics and ontology, critical aspects regarding clinical practice. A wise approach to the problem would consist of integrating inclusively powerful data-driven methods with model-driven ones into explainable ML (XML) [22], integrating the two approaches towards exploring the best of both worlds instead of mutual auto-exclusion.

The study has some limitations to be considered. As mentioned before, the number of works included in the present study is a small subset of what has been published recently. This fact being true, it must be mentioned that the number of publications meeting the strict criteria adopted is low, as many recent publications do not fulfill the selection criteria, and including or even mentioning them would be distracting the reader through a jungle of publications of minor relevance. Another limitation is the controversial issue of subjectivity. The authors are aware that this is a major question, but as mentioned before, criticism and discussion are the ways to scientific advancements. That is why the term critical has been included in the title.

5. Conclusions

The way that ML methods are being applied to laryngeal pathology assessment by many current publications trying to classify "normal" vs. "pathological" seems to be at least somehow problematic. Besides offering a shallow view of a priori knowledge on the topic, many of them stand on rather subjective foundations that are not well supported by statistical validation. This large number of approaches obscures the truly interesting research publications from the reader, much in the same way as when one cannot see the wood for the trees. With this aim in mind, the present paper has tried to establish certain criteria to examine the quality and relevance that current publications in the field should meet to be considered meriting the attention of the research community on this issue. The study analyzed 26 publications after a selection process, and 15 of them were found to be MD inspired, whereas seven others showed a marked DD character. Four shared both profiles. In eighteen of them, results were availed by some statistical methodology. Clinical relevance issues were commented in 14 of them.

The following list enumerates the criteria found to be fulfilled by possible publications addressing the issue of laryngeal pathology assessment:

- Trustworthiness
- Statistical Relevance
- Originality
- Clinical Explainability
- Self-criticism

Another relevant issue is that of methodological experimental design to preserve the reliability and trustworthiness of results. This is the list of requirements to be met:

- Database size and balance
- Gender separation
- Correct speech database production
- Careful feature extraction
- Problem orientation definition
- Clinical semantic power and acceptance
- Ethical implications
- Information integrity preservation
- Linguistic issues
- Relevant metadata
- Adequate model selection and description
- Discarding irrelevant subjective improvements
- Preservation of semantics and future hypothesis casting

One of the most relevant from the list above is clinical explainability, which deserves special attention for its definition and applicability. Quoting Górriz et al. [94] "... it is necessary to find additional mechanisms that explain the reasoning behind how conclusions have been reached, to achieve systems that are reliable and, therefore, easier to deploy in sensitive areas, such as health or security, in which machines interact with humans... This explanation is already a recognized right for the European Union". Consequently, as a final remark, it must be mentioned that EE and similar DL approaches maximizing the power of deep ML have to seriously consider the preservation of clinical explainability and semantics of results, being especially careful not to spread dubious expectations on using terms related with explanation-less automatic diagnosis.

This critical appraisal concludes by suggesting possible potential strategies to overcome some of the lacks and wants to be detected in the publications examined, to help researchers in the field look for real advances in ML-assisted voice quality assessment.

Author Contributions: Conceptualization, A.G.-R. and D.P.-A.; Formal analysis, A.G.-R.; Supervision, A.Á.-M.; Writing—original draft, P.G.-V.; Writing—review & editing, D.P.-A. and V.R.-B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by King Juan Carlos University Project "Complex morphing and presentation attack detection in uncontrolled high-security scenarios" (COMPAD—M2606).

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Ethics Committee of Universidad Politécnica de Madrid (protocol MonParLoc of 21 May 2018).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Excluded, data used were taken from public sources.

Acknowledgments: Andrés Gómez Rodellar holds a scholarship from the Medical Research Council Doctoral Training Programme in the Usher's Institute (University of Edinburgh Medical School).

Conflicts of Interest: The authors declare no conflict of interest.

References and Notes

1. Sahin, S.; Tolun, M.R.; Hassanpour, R. Hybrid expert systems: A survey of current approaches and applications. *Expert Syst. Appl.* **2012**, *39*, 4609–4617. [CrossRef]
2. Keener, J.; Sneyd, J. *Mathematical Physiology: II: Systems Physiology*; Springer: Berlin/Heidelberg, Germany, 2009.
3. Titze, I. *Principles of Voice Production*; Prentice-Hall: Hoboken, NJ, USA, 1994.
4. Hastie, T.; Tibshirani, R.; Friedman, J.H.; Friedman, J.H. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*; Springer: Berlin/Heidelberg, Germany, 2009.
5. Stark, J.; Hardy, K. Chaos: Useful at last? *Science* **2003**, *301*, 1192–1193. [CrossRef] [PubMed]
6. Narendra, N.P.; Alku, P. Glottal source information for pathological voice detection. *IEEE Access* **2020**, *8*, 67745–67755. [CrossRef]
7. Fant, G. The source filter concept in voice production. *STL-QPSR* **1981**, *1*, 21–37.
8. Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Giannotti, F.; Pedreschi, D. A Survey of Methods for Explaining Black Box Models. *ACM Comput. Surv.* **2019**, *51*, 1–42. [CrossRef]
9. Hlavnička, J.; Čmejla, R.; Tykalová, T.; Šonka, K.; Růžička, E.; Rusz, J. Automated analysis of connected speech reveals early biomarkers of Parkinson's disease in patients with rapid eye movement sleep behaviour disorder. *Sci. Rep.* **2017**, *7*, 12. [CrossRef]
10. ASHA. Voice Disorders. Available online: https://www.asha.org/Practice-Portal/Clinical-Topics/Voice-Disorders/ (accessed on 13 July 2022).
11. Schmid, L.; Gerharz, A.; Groll, A.; Pauly, M. Machine Learning for Multi-Output Regression: When should a holistic multivariate approach be preferred over separate univariate ones? *arXiv* **2022**, arXiv:2201.05340.
12. Saarbrücken Voice Database. Available online: http://www.stimmdatenbank.coli.uni-saarland.de/help_en.php4 (accessed on 21 October 2022).
13. Amato, F.; Borzì, L.; Olmo, G.; Orozco-Arroyave, J.R. An algorithm for Parkinson's disease speech classification based on isolated words analysis. *Health Inf. Sci. Syst.* **2021**, *9*, 32. [CrossRef]
14. Godino-Llorente, J.I.; Gómez-Vilda, P.; Cruz-Roldán, F.; Blanco-Velasco, M.; Fraile, R. Pathological Likelihood Index as a Measurement of the Degree of Voice Normality and Perceived Hoarseness. *J. Voice* **2010**, *24*, 667–677. [CrossRef]
15. MEEI Database, Massachusetts Eye and Ear Infirmary Voice and Speech Lab, Boston, MA. &. KayPENTAX, Kay Elemetrics Disordered Voice Database, Model 4337. Kay Elemetrics, Lincoln Park, NJ, USA. 1996–2005.
16. Holzinger, A.; Langs, G.; Denk, H.; Zatloukal, K.; Müller, H. Causability and explainability of artificial intelligence in medicine. *WIREs Data Min. Knowl. Discov.* **2019**, *9*, e1312. [CrossRef]
17. Mekyska, J.; Janousova, E.; Gomez-Vilda, P.; Smekal, Z.; Rektorova, I.; Eliasova, I.; Kostalova, M.; Mrackova, M.; Alonso-Hernandez, J.B.; Faundez-Zanuy, M.; et al. Robust and complex approach of pathological speech signal analysis. *Neurocomputing* **2015**, *167*, 94–111. [CrossRef]
18. Travieso, C.M.; Alonso, J.B.; Orozco-Arroyave, J.; Vargas-Bonilla, J.; Nöth, E.; Ravelo-García, A.G. Detection of different voice diseases based on the nonlinear characterization of speech signals. *Expert Syst. Appl.* **2017**, *82*, 184–195. [CrossRef]
19. Dejonckere, P.H.; Bradley, P.; Clemente, P.; Cornut, G.; Crevier-Buchman, L.; Friedrich, G.; Van De Heyning, P.; Remacle, M.; Woisard, V. A basic protocol for functional assessment of voice pathology, especially for investigating the efficacy of (phonosurgical) treatments and evaluating new assessment techniques. *Eur. Arch. Oto-Rhino-Laryngol.* **2001**, *258*, 77–82. [CrossRef] [PubMed]
20. Al-Nasheri, A.; Muhammad, G.; Alsulaiman, M.; Ali, Z.; Malki, K.H.; Mesallam, T.A.; Ibrahim, M.F. Voice Pathology Detection and Classification Using Auto-Correlation and Entropy Features in Different Frequency Regions. *IEEE Access* **2017**, *6*, 6961–6974. [CrossRef]
21. Magner, L.N.; Kim, O.J. *A History of Medicine*; CRC Press: Boca Raton, FL, USA, 2017.
22. Tjoa, E.; Guan, C. A survey on explainable artificial intelligence (XAI): Toward medical XAI. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *32*, 4793–4813. [CrossRef]

23. Madiega, T.A. EU Guidelines on Ethics in Artificial Intelligence: Context and Implementation. EPRS: European Parliamentary Research Service. 2019. Available online: https://policycommons.net/artifacts/1337743/eu-guidelines-on-ethics-in-artificial-intelligence/1945725/ (accessed on 13 July 2022).
24. Li, X.; Jiang, Y.; Li, M.; Yin, S. Lightweight Attention Convolutional Neural Network for Retinal Vessel Image Segmentation. *IEEE Trans. Ind. Inf.* **2020**, *17*, 1958–1967. [CrossRef]
25. Jiang, Y.; Li, X.; Luo, H.; Yin, S.; Kaynak, O. Quo vadis artificial intelligence? *Discov. Artif. Intell.* **2022**, *2*, 4. [CrossRef]
26. Volovici, V.; Syn, N.L.; Ercole, A.; Zhao, J.J.; Liu, N. Steps to avoid overuse and misuse of machine learning in clinical research. *Nat. Med.* **2022**, *28*, 1996–1999. [CrossRef]
27. Uloza, V.; Verikas, A.; Bacauskiene, M.; Gelzinis, A.; Pribuisiene, R.; Kaseta, M.; Saferis, V. Categorizing Normal and Pathological Voices: Automated and Perceptual Categorization. *J. Voice* **2011**, *25*, 700–708. [CrossRef]
28. Verikas, A.; Gelzinis, A.; Vaiciukynas, E.; Bacauskiene, M.; Minelga, J.; Hållander, M.; Uloza, V.; Padervinskis, E. Data dependent random forest applied to screening for laryngeal disorders through analysis of sustained phonation: Acoustic versus contact microphone. *Med. Eng. Phys.* **2015**, *37*, 210–218. [CrossRef]
29. Martins, R.H.G.; Amaral, H.A.D.; Tavares, E.L.M.; Martins, M.G.; Gonçalves, T.M.; Dias, N.H. Voice Disorders: Etiology and Diagnosis. *J. Voice* **2016**, *30*, 761.e1–761.e9. [CrossRef] [PubMed]
30. Topol, E.J. High-performance medicine: The convergence of human and artificial intelligence. *Nat. Med.* **2019**, *25*, 44–56. [CrossRef] [PubMed]
31. Harar, P.; Galaz, Z.; Alonso-Hernandez, J.B.; Mekyska, J.; Burget, R.; Smekal, Z. Towards robust voice pathology detection. *Neural Comput. Appl.* **2020**, *32*, 15747–15757. [CrossRef]
32. Saibene, A.; Assale, M.; Giltri, M. Expert systems: Definitions, advantages and issues in medical field applications. *Expert Syst. Appl.* **2021**, *177*, 114900. [CrossRef]
33. Heckerman, D.E.; Shortliffe, E.H. From certainty factors to belief networks. *Artif. Intell. Med.* **1992**, *4*, 35–52. [CrossRef]
34. Rosenblatt, F. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychol. Rev.* **1958**, *65*, 386–408. [CrossRef] [PubMed]
35. Minsky, M.L. Logical versus analogical or symbolic versus connectionist or neat versus scruffy. *AI Mag.* **1991**, *12*, 34. [CrossRef]
36. Titze, I.R. Current topics in voice production mechanisms. *Acta Oto-Laryngol.* **1993**, *113*, 421–427. [CrossRef] [PubMed]
37. Vergin, R.; O'Shaughnessy, D.; Farhat, A. Generalized mel frequency cepstral coefficients for large-vocabulary speaker-independent continuous-speech recognition. *IEEE Trans. Speech Audio Process.* **1999**, *7*, 525–532. [CrossRef]
38. Tsanas, A. Relevance, redundancy and complementarity trade-off (RRCT): A generic, efficient, robust feature selection tool. *Gene Expr. Patterns* **2022**, *3*, 100471. [CrossRef]
39. Breiman, L. Statistical Modeling: The two cultures. *Statist. Sci.* **2001**, *16*, 199–231. [CrossRef]
40. Forsyth, D. *Applied Machine Learning*; Springer: Berlin/Heidelberg, Germany, 2019. [CrossRef]
41. Kelly, C.J.; Karthikesalingam, A.; Suleyman, M.; Corrado, G.; King, D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med.* **2019**, *17*, 195. [CrossRef] [PubMed]
42. Ashri, R. Building AI Software: Data-Driven vs. Model-Driven AI and Why We Need an AI-Specific Software (Issues Brief). 2017. Available online: https://hackernoon.com/building-ai-softwaredata-driven-vs-model-driven-ai-and-why-we-need-an-specific-software-640f74aaf78f (accessed on 14 July 2022).
43. Maruyama, Y. Symbolic and statistical theories of cognition: Towards integrated artificial intelligence. In *International Conference on Software Engineering and Formal Methods*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 129–146. [CrossRef]
44. Taroni, F.; Bozza, S.; Biedermann, A.; Garbolino, P.; Aitken, C. *Data Analysis in Forensic Science: A Bayesian Decision Perspective*; John Wiley & Sons: Hoboken, NJ, USA, 2010.
45. Abitbol, J.; Abitbol, P.; Abitbol, B. Sex hormones and the female voice. *J. Voice* **1999**, *13*, 424–446. [CrossRef]
46. Inamoto, Y.; Saitoh, E.; Okada, S.; Kagaya, H.; Shibata, S.; Baba, M.; Onogi, K.; Hashimoto, S.; Katada, K.; Wattanapan, P.; et al. Anatomy of the larynx and pharynx: Effects of age, gender and height revealed by multidetector computed tomography. *J. Oral Rehabil.* **2015**, *42*, 670–677. [CrossRef] [PubMed]
47. Davatz, G.C.; Yamasaki, R.; Hachiya, A.; Tsuji, D.H.; Montagnoli, A.N. Source and Filter Acoustic Measures of Young, Middle-Aged and Elderly Adults for Application in Vowel Synthesis. *J. Voice* **2021**, *in press*. [CrossRef] [PubMed]
48. Whiteside, S.P.; Hodgson, C. Some acoustic characteristics in the voices of 6- to 10-year-old children and adults: A comparative sex and developmental perspective. *Logop. Phoniatr. Vocol.* **2000**, *25*, 122–132. [CrossRef] [PubMed]
49. Gómez-Vilda, P.; Fernández-Baillo, R.; Rodellar-Biarge, V.; Lluis, V.N.; Álvarez-Marquina, A.; Mazaira-Fernández, L.M.; Martínez-Olalla, R.; Godino-Llorente, J.I. Glottal Source biometrical signature for voice pathology detection. *Speech Commun.* **2009**, *51*, 759–781. [CrossRef]
50. Cirillo, D.; Catuara-Solarz, S.; Morey, C.; Guney, E.; Subirats, L.; Mellino, S.; Gigante, A.; Valencia, A.; Rementeria, M.J.; Chadha, A.S.; et al. Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare. *NPJ Digit. Med.* **2020**, *3*, 81. [CrossRef]
51. Mellino, S.; Morey, C.; Rohner, C. Biases in digital health measures. In *Sex and Gender Bias in Technology and Artificial Intelligence*; Elsevier: Amsterdam, The Netherlands, 2022; pp. 95–112. [CrossRef]

52. Bouckaert, R.R.; Frank, E. Evaluating the replicability of significance tests for comparing learning algorithms. In Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, 26–28 May 2004; Springer: Berlin/Heidelberg, Germany, 2004; pp. 3–12. [CrossRef]
53. Demšar, J. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.* **2006**, *7*, 1–30.
54. Hand, D.J.; Till, R.J. A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems. *Mach. Learn.* **2001**, *45*, 171–186. [CrossRef]
55. Waibel, A.; Hanazawa, T.; Hinton, G.; Shikano, K.; Lang, K. Phoneme recognition using time-delay neural networks. *IEEE Trans. Acoust. Speech Signal Process.* **1989**, *37*, 328–339. [CrossRef]
56. Al-Dhief, F.T.; Baki, M.M.; Latiff, N.M.A.; Malik, N.N.N.A.; Salim, N.S.; Albader, M.A.A.; Mahyuddin, N.M.; Mohammed, M.A. Voice Pathology Detection and Classification by Adopting Online Sequential Extreme Learning Machine. *IEEE Access* **2021**, *9*, 77293–77306. [CrossRef]
57. Albtoush, A.; Fernández-Delgado, M.; Cernadas, E.; Barro, S. Quick extreme learning machine for large-scale classification. *Neural Comput. Appl.* **2022**, *34*, 5923–5938. [CrossRef]
58. Hammami, I.; Salhi, L.; Labidi, S. Voice Pathologies Classification and Detection Using EMD-DWT Analysis Based on Higher Order Statistic Features. *IRBM* **2020**, *41*, 161–171. [CrossRef]
59. Vaziri, G.; Almasganj, F.; Behroozmand, R. Pathological assessment of patients' speech signals using nonlinear dynamical analysis. *Comput. Biol. Med.* **2010**, *40*, 54–63. [CrossRef] [PubMed]
60. Tennenholtz, G.; Zahavy, T.; Mannor, S. Train on validation: Squeezing the data lemon. *arXiv* **2018**, arXiv:1802.05846. [CrossRef]
61. Arias-Londoño, J.D.; Godino-Llorente, J.I.; Sáenz-Lechón, N.; Osma-Ruiz, V.; Castellanos-Domínguez, G. Automatic Detection of Pathological Voices Using Complexity Measures, Noise Parameters, and Mel-Cepstral Coefficients. *IEEE Trans. Biomed. Eng.* **2010**, *58*, 370–379. [CrossRef]
62. Zhao, Y.; Foong, L.K. Predicting electrical power output of combined cycle power plants using a novel artificial neural network optimized by electrostatic discharge algorithm. *Measurement* **2022**, *198*, 111405. [CrossRef]
63. Zhao, Y.; Wang, Z. Subset simulation with adaptable intermediate failure probability for rogust reliability analysis: And unsupervised learning-based approach. *Struct. Multidiscip. Optim.* **2022**, *65*, 172. [CrossRef]
64. Ahuja, A.S. The impact of artificial intelligence in medicine on the future role of the physician. *PeerJ* **2019**, *7*, e7702. [CrossRef]
65. Lee, J.-Y. Experimental evaluation of deep learning methods for an intelligent pathological voice detection system using the Saarbruecken voice database. *Appl. Sci.* **2021**, *11*, 7149. [CrossRef]
66. Gómez-García, J.A.; Moro-Velázquez, L.; Godino-Llorente, J.I. On the design of automatic voice condition analysis systems. Part I: Review of concepts and an insight to the state of the art. *Biomed. Signal Process. Control* **2019**, *51*, 181–199. [CrossRef]
67. Islam, R.; Tarique, M.; Abdel-Raheem, E. A survey on signal processing based pathological voice detection techniques. *IEEE Access* **2020**, *8*, 66749–66776. [CrossRef]
68. Hegde, S.; Shetty, S.; Rai, S.; Dodderi, T. A Survey on Machine Learning Approaches for Automatic Detection of Voice Disorders. *J. Voice* **2019**, *33*, 947.e11–947.e33. [CrossRef] [PubMed]
69. Orozco-Arroyave, J.R.; Belalcazar-Bolanos, E.A.; Arias-Londono, J.D.; Vargas-Bonilla, J.F.; Skodda, S.; Rusz, J.; Daqrouq, K.; Honig, F.; Noth, E. Characterization Methods for the Detection of Multiple Voice Disorders: Neurological, Functional, and Laryngeal Diseases. *IEEE J. Biomed. Health Inf.* **2015**, *19*, 1820–1828. [CrossRef] [PubMed]
70. Akbari, A.; Arjmandi, M.K. An efficient voice pathology classification scheme based on applying multi-layer linear discriminant analysis to wavelet packet-based features. *Biomed. Signal Process. Control* **2014**, *10*, 209–223. [CrossRef]
71. Moro-Velázquez, L.; Gómez-García, J.A.; Godino-Llorente, J.I. Voice Pathology Detection Using Modulation Spectrum-Optimized Metrics. *Front. Bioeng. Biotechnol.* **2016**, *4*, 1. [CrossRef]
72. Verde, L.; de Pietro, G.; Sannino, G. Voice disorder identification by using machine learning techniques. *IEEE Access* **2018**, *6*, 16246–16255. [CrossRef]
73. Pützer, M.; Wokurek, W. Electroglottographic and Acoustic Parametrization of Phonatory Quality Provide Voice Profiles of Pathological Speakers. *J. Voice* **2021**, *in press*. [CrossRef]
74. Hemmerling, D.; Skalski, A.; Gajda, J. Voice data mining for laryngeal pathology assessment. *Comput. Biol. Med.* **2016**, *69*, 270–276. [CrossRef]
75. Barreira, R.R.A.; Ling, L.L. Kullback–Leibler divergence and sample skewness for pathological voice quality assessment. *Biomed. Signal Process. Control* **2020**, *57*, 101697. [CrossRef]
76. Omeroglu, A.N.; Mohammed, H.M.A.; Oral, E.A. Multi-modal voice pathology detection architecture based on deep and handcrafted feature fusion. *Eng. Sci. Technol. Int. J.* **2022**, *36*, 101148. [CrossRef]
77. Ding, H.; Gu, Z.; Dai, P.; Zhou, Z.; Wang, L.; Wu, X. Deep connected attention (DCA) ResNet for robust voice pathology detection and classification. *Biomed. Signal Process. Control* **2021**, *70*, 102973. [CrossRef]
78. Kadiri, S.R.; Alku, P. Analysis and detection of pathological voice using glottal source features. *IEEE J. Sel. Top. Signal Process.* **2019**, *14*, 367–379. [CrossRef]
79. Chen, L.; Chen, J. Deep neural network for automatic classification of pathological voice signals. *J. Voice* **2020**, *36*, 288.e15–288.e24. [CrossRef]

80. Mohammed, M.A.; Abdulkareem, K.H.; Mostafa, S.A.; Ghani, M.K.A.; Maashi, M.S.; Garcia-Zapirain, B.; Oleagordia, I.; AlHakami, H.; Al-Dhief, F.T. Voice Pathology Detection and Classification Using Convolutional Neural Network Model. *Appl. Sci.* **2020**, *10*, 3723. [CrossRef]
81. Wu, Y.; Zhou, C.; Fan, Z.; Wu, D.; Zhang, X.; Tao, Z. Investigation and Evaluation of Glottal Flow Waveform for Voice Pathology Detection. *IEEE Access* **2020**, *9*, 30–44. [CrossRef]
82. Zhou, C.; Wu, Y.; Fan, Z.; Zhang, X.; Wu, D.; Tao, Z. Gammatone spectral latitude features extraction for pathological voice detection and classification. *Appl. Acoust.* **2021**, *185*, 108417. [CrossRef]
83. Olson, R.S.; Moore, J.H. TPOT: A Tree-Based Pipeline Optimization Tool for Automating Machine Learning. In Proceedings of the Workshop on Automatic Machine Learning, New York, NY, USA, 24 June 2016; pp. 66–74. Available online: http://proceedings.mlr.press/v64/olson_tpot_2016.pdf (accessed on 19 September 2022).
84. LeDell, E.; Poirier, S. H_2O automl: Scalable automatic machine learning. In Proceedings of the AutoML Workshop at ICML, online, 17–18 July 2020; Volume 2020.
85. Creswell, A.; White, T.; Dumoulin, V.; Arulkumaran, K.; Sengupta, B.; Bharath, A.A. Generative Adversarial Networks: An Overview. *IEEE Signal Process. Mag.* **2018**, *35*, 53–65. [CrossRef]
86. El Amri, W.Z.; Reinhart, F.; Schenck, W. Open set task augmentation facilitates generalization of deep neural networks trained on small data sets. *Neural Comput. Appl.* **2022**, *34*, 6067–6083. [CrossRef]
87. Patel, R.R.; Awan, S.N.; Barkmeier-Kraemer, J.; Courey, M.; Deliyski, D.; Eadie, T.; Paul, D.; Svec, J.G.; Hillman, R. Recommended Protocols for Instrumental Assessment of Voice: American Speech-Language-Hearing Association Expert Panel to Develop a Protocol for Instrumental Assessment of Vocal Function. *Am. J. Speech-Language Pathol.* **2018**, *27*, 887–905. [CrossRef] [PubMed]
88. James, G.; Witten, D.; Hastie, T.; Tibshirani, R. *An Introduction to Statistical Learning*; Springer: Berlin/Heidelberg, Germany, 2013; Volume 112. [CrossRef]
89. Li, J.; Cheng, K.; Wang, S.; Morstatter, F.; Trevino, R.P.; Tang, J.; Liu, H. Feature Selection. *ACM Comput. Surv.* **2018**, *50*, 1–45. [CrossRef]
90. Urbanowicz, R.J.; Meeker, M.; La Cava, W.; Olson, R.S.; Moore, J.H. Relief-based feature selection: Introduction and review. *J. Biomed. Inf.* **2018**, *85*, 189–203. [CrossRef]
91. Bernard, M.; Poli, M.; Karadayi, J.; Dupoux, E. Shennong: A Python toolbox for audio speech features extraction. *arXiv* **2021**, arXiv:2112.05555.
92. Wang, K.; Gou, C.; Duan, Y.; Lin, Y.; Zheng, X.; Wang, F.-Y. Generative adversarial networks: Introduction and outlook. *IEEE/CAA J. Autom. Sin.* **2017**, *4*, 588–598. [CrossRef]
93. Oyelade, O.N.; Ezugwu, A.E.; Almutairi, M.S.; Saha, A.K.; Abualigah, L.; Chiroma, H. A generative adversarial network for synthetization of regions of interest based on digital mammograms. *Sci. Rep.* **2022**, *12*, 6166. [CrossRef]
94. Górriz, J.M.; Ramírez, J.; Ortíz, A.; Martínez-Murcia, F.J.; Segovia, F.; Suckling, J.; Leming, M.; Zhang, Y.-D.; Álvarez-Sánchez, J.R.; Bologna, G.; et al. Artificial intelligence within the interplay between natural and artificial computation: Advances in data science, trends and applications. *Neurocomputing* **2020**, *410*, 237–270. [CrossRef]