Non-invasive AI-powered Diagnostics: The case of Voice-Disorder Detection - Vision paper

Gabriele Ciravegna[1], Alkis Koudounas[1], Marco Fantini[2], Tania Cerquitelli[1], Elena Baralis[1], Erika Crosetti[3] and Giovanni Succo[3]

[1]Politecnico di Torino, Corso Duca degli Abruzzi, Turin, Italy
[2]ENT Unit, San Feliciano Hospital, Rome, Italy
[3]ENT Clinic - Head and Neck Cancer Unit, San Giovanni Bosco Hospital, Turin Italy

Abstract

This paper proposes a novel pipeline for non-invasive diagnosis and monitoring in healthcare, leveraging artificial intelligence (AI). The pipeline allows individuals to record various health data using everyday devices and analyze it via AI algorithms on a cloud-based platform. Experimental results on voice disorder detection demonstrate the effectiveness of the proposed approach when compared to existing solutions. Additionally, we discuss the positive impact of the pipeline on diagnosis, prognosis, and monitoring, emphasizing its non-invasive nature. Overall, we think the proposed pipeline might contribute to advancing AI-driven healthcare solutions with implications for global healthcare delivery.

Keywords

Artificial Intelligence, Non-invasive diagnostics, Voice disorder recognition, Voice analysis

1. Introduction

Artificial Intelligence (AI) is increasingly integrated into healthcare, offering opportunities to improve diagnostics, treatment, and patient care. Through machine learning algorithms, AI systems can analyze medical data, providing clinical decision support and personalized treatment options [1]. Non-invasive diagnostics is a critical aspect of modern healthcare delivery, offering patients a less intrusive and more comfortable care experience. This holds for both the diagnostic and the monitoring processes, improving the overall quality of life for patients [2]. Beyond enhancing patient comfort, these methods also improve accessibility to healthcare services, particularly for underserved populations or those with limited access to specialized medical facilities.

We propose an AI-based framework for non-invasive diagnostics that integrates various data modalities, including voice recordings, self-pictures, typing patterns, ECG readings, and sleep analysis, among others. These diverse sources of data are collected through everyday devices such as computers, smartphones, and smartwatches, enabling convenient and continuous monitoring of individual health metrics. Subsequently, these data are fed into cloud-based applications where advanced AI algorithms analyze them, identifying patterns that may be indicative of underlying health issues. Also, the integration of multiple data sources allows for a holistic understanding of an individual's health, facilitating early detection and personalized intervention strategies.

The envisioned framework has the potential to improve healthcare delivery as well as have economic and technological impacts. First, by enabling large scale screening, it may increase early detection, allowing for timely intervention and improved treatment outcomes. Second, by analyzing the evolution of patient data in time, healthcare providers may tailor interventions, optimizing treatment efficacy and patient satisfaction. From an economic perspective, the improved efficiency of diagnosis and treatments may reduce overall healthcare costs. Finally, the developed models may be employed or fine-tuned in related data-scarce contexts.

As part of our investigation, we conducted a preliminary study on voice disorder detection, a key aspect of non-invasive diagnostics. We trained a deep learning model for analyzing voice recordings that achieves very high precision in detecting the presence of pathology and accurately identifies the type of pathology. We employed a transformer model [3], trained end-to-end (E2E) directly on the raw data, outperforming traditional methods such as convolutional neural networks (CNN) trained on frequency-transformed data. These preliminary results demonstrate the potential of our approach in enabling accurate and efficient non-invasive diagnostics for voice disorders.

The paper begins with an introduction to AI and its medical applications (Section 2). It then outlines the proposed idea, and it addresses foreseen challenges (Section 3). We then present the preliminary results on a voice disorder detection case in Section 4, and conclude with a discussion on the framework's impact (Section 5).

Published in the Proceedings of the Workshops of the EDBT/ICDT 2024 Joint Conference (March 25-28, 2024), Paestum, Italy

gabriele.ciravegna@polito.it (G. Ciravegna)

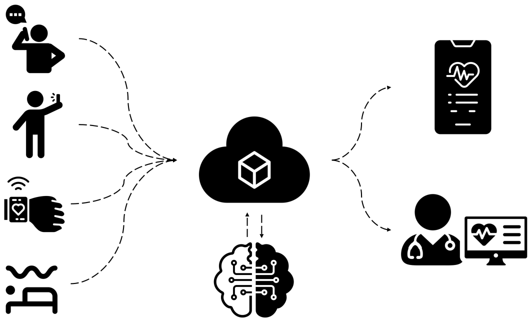

Figure 1: Outline of the proposed pipeline. A user records different types of data, e.g., pictures, voice, texts, heart monitors, and sleep conditions. These data are uploaded to the cloud and processed by an artificial intelligence method. The results are visualized on a user-controlled application but can also be shared with a remote doctor who can require further exams.

2. Background

Deep Learning and Transformers Deep Learning is a robust method for uncovering patterns and insights from extensive datasets [4]. Unlike conventional machine learning approaches, Deep Learning models learn directly from raw data in an E2E manner, without the need for manual feature engineering. Transformer models are a prominent class of Deep Learning architectures [3], demonstrating the effectiveness of this approach in analyzing sequential and multimodal data [5, 6]. Unlike traditional recurrent and convolutional neural networks, Transformers can capture long-range dependencies and preserve contextual information over extended sequences, thanks to their self-attention mechanisms. This characteristic enables Transformers to process raw multimodal sequential data, such as text or time-series data, as well as images, making them highly adaptable for various applications, including medical diagnostics [7, 8].

DL for Medicine Deep Learning has emerged as a transformative technology in various healthcare applications. First, DL can analyze electronic health records [9], enabling personalized treatment recommendations and predictive analytics for patient outcomes. In medical imaging [10], DL models can interpret radiological images, including X-rays, CT scans, and MRI images, with levels of accuracy comparable to or even surpassing that of human experts. These models have been utilized for tasks such as disease classification, lesion detection, and tumor segmentation, the latter also performed in real-time during surgery. In drug discovery, AI algorithms have been shown capable of developing novel drugs through accurate predictions of protein structures [11]. Finally, DL models have also shown increased capability in handling multi-omics and multi-modal data [12].

3. AI for non-invasive medicine

In this paper, we propose a framework for analyzing an individual's health condition in time through non-invasive AI-powered diagnostics. Our proposed pipeline embodies a user-centric approach, empowering individuals to actively participate in their healthcare journey. From a medical point of view, we consider the analysis of all pathologies detectable by human experts through non-invasive diagnosis. Drawing parallels with advancements in other domains, we argue that AI models can replicate and potentially surpass human diagnostic accuracy also in this domain. This solution may enable large-scale screening through simple but accurate analysis of the patient data collected remotely in a non-invasive way.

Data collection In Figure 1 we report a visualization of the proposed framework. Leveraging everyday devices such as smartphones, laptops, or wearables, users can effortlessly record diverse types of data. Clearly, each type of device enables the collection of different data types. Laptops and smartphones allow collecting typing and click patterns, texts, voice recordings, and pictures (the last two particularly through smartphones). Wearables, on the other side, allow (and are already employed for) monitoring heart rate, blood pressure, athletic performance and sleep conditions, to cite a few.

Data storage analysis Upon collection, the data are uploaded to a cloud-based database and analyzed by means of a single advanced artificial intelligence model. The employment of the aforementioned transformer models accommodate diverse data modalities and sources. By integrating multi-modal data processing capabilities, our framework aims to capture a holistic view of an individual's health status, enabling comprehensive and personalized diagnostics. Moreover, the collection of continuous health data enable taking into consideration in-time evolution and predictive modelling. The model may also evolve and improve over time by means of the new data.

Result visualization The processed results are then presented to users through an intuitive and user-controlled application interface. Through this interface, users gain valuable insights into their health metrics, facilitating informed and proactive decisions regarding their healthcare. Furthermore, a user can transmit the processed data and diagnostic results to their remote healthcare professionals, allowing them to make informed clinical decisions and make timely interventions. Additionally, healthcare providers can utilize the data for population-level health monitoring and epidemiological studies, facilitating the identification of emerging health trends and proactive public health interventions.

3.1. Challenges

The effectiveness of the proposed framework relies also on our ability to address and resolve a number of both technical and non-technical issues.

3.1.1. Data quality

The collection of data through personal devices introduces the risk of data noise, which encompasses various factors such as sensor inaccuracies, environmental interference, and incorrect user input. Sensor inaccuracies may lead to erroneous measurements, while environmental interference, such as background noise or lighting conditions, can distort the recorded data. Additionally, users may input incorrect data, such as taking pictures of the wrong part of their body, which can further exacerbate data noise and impact the analysis model. These challenges can hinder the model's ability to accurately interpret and analyze the data, potentially leading to erroneous conclusions and suboptimal performance.

Research direction: data augmentation + input data checking To address the challenges posed by sensor inaccuracies and environmental interferences, a robust data augmentation pipeline can be implemented to mitigate these sources of noise. By incorporating various data augmentation techniques such as noise injection, signal filtering, and data synthesis, the pipeline can generate diverse and representative training data that encapsulates the variability present in real-world scenarios [13]. Specific preprocessing tailored to each data modality can help normalize and enhance the quality of the collected data. Furthermore, to mitigate the issue of incorrect user input, a dedicated model can be employed to validate the accuracy of the collected data. This model can check each type of collected data, such as images or sensor readings, to verify their correctness and flag any discrepancies.

3.1.2. Privacy preservation

The second problem arises from privacy concerns associated with collecting sensitive data and transmitting it to a cloud-based database. Individuals may be hesitant to share sensitive information due to concerns about data security and privacy breaches [14]. Transmitting such data to a cloud-based database further exacerbates these concerns, as it involves relinquishing control over personal information to third-party service providers. Moreover, regulatory compliance [15] impose stringent requirements on the handling of personal health information, adding complexity to the data collection process.

Research direction: a federated learning approach The privacy issue can be effectively addressed through the adoption of a federated approach. Federated learning is an area of machine learning studying how to coordinate and distribute training over several models without exchanging raw data. The goal is to mitigate privacy concerns associated with transmitting sensitive data to a centralized cloud-based database. Furthermore, personalized models generated through federated learning may exhibit higher diagnostic accuracy, as they are trained on data specific to each user's health profile.

3.1.3. Human understanding & trustworthiness

The third issue concerns the explainability challenge associated with employing deep learning-based black box models for predictions. While these models offer impressive performance, their inherent complexity makes them opaque and difficult to interpret. This lack of explainability presents challenges for clinicians and end-users who require insights into the model's decision-making process [16, 17, 18, 19]. Without transparent explanations, stakeholders and regulatory institutions may hesitate to trust the diagnostic and monitoring framework.

Research direction: Concept-based XAI models A possible solution is represented by the employment of eXplainable AI (XAI) algorithms that can shed light on the decisions of the models [20, 21, 22]. Particularly, Concept-based XAI models offer intrinsic interpretability by mapping raw data to interpretable high-level concepts before making class predictions, offering insights into the model's decision-making process [23, 24, 25]. This approach not only addresses the explainability challenge but also fosters trust and confidence in the diagnostic and monitoring outcomes produced by the system.

4. The case of voice disorder detection

Vocal disorders are prevalent pathologies affecting a significant portion of the population and exerting a substantial impact on patients' quality of life [26, 27, 28, 29]. These disorders may originate from various causes, including both benign and malignant conditions, and neurodegenerative disorders [30, 31, 32]. Diagnosis often relies on clinicians' auditory assessments of patients' voices, highlighting the critical need for accurate and timely detection. Here, a DL model is used to analyze the raw recordings and automatically detect patterns indicative of vocal disorders and distinguish between various pathologies, including nodules, polyps, cysts, spasmodic dysphonia or vocal cord paralysis.

Preliminary experiments We demonstrate significant advancements over prior attempts in voice disorder detection using AI models [33, 34][1]. As reported in Figure 2, our approach achieves notable improvements in

[1] We re-implemented these models for a fair comparison.

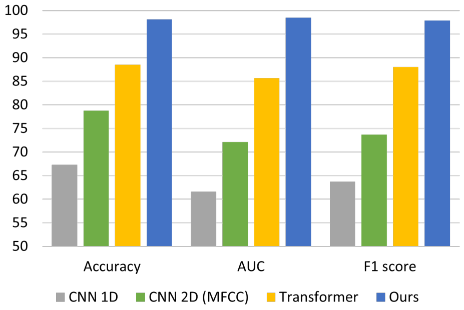

Figure 2: Comparative analysis of the test model performance in distinguishing Healthy individuals from those afflicted with a pathological condition.

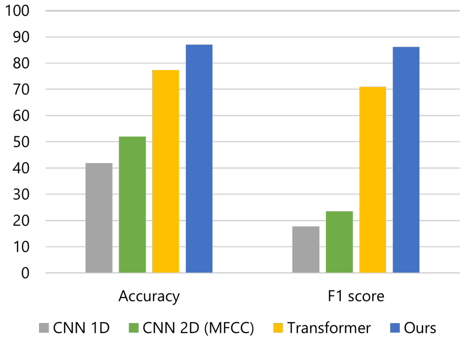

Figure 3: Comparative analysis of test model performance in identifying the macro pathology that afflicts individuals.

accuracy, up to +30%, primarily attributed to the utilization of a Transformer model rather than a Convolutional Neural Network (CNN). The Transformer's inherent ability to process raw time-series data E2E without any time-frequency preprocessing offers distinct advantages in analyzing voice data [35]. Additionally, we introduce a robust data augmentation pipeline and consider both vowel and sentence-based recordings, further enhancing performance by up to +10% compared to the employment of a standard Transformer model only. As reported in Figure 3, similar considerations hold also for the classification of the macro-category pathology. These enhancements underscore the efficacy of our approach in the diagnosing vocal disorders.

5. Discussion

Overall, the proposed pipeline holds promise for improving healthcare delivery, leveraging AI to enable non-invasive diagnostics and monitoring. The medical team involved in the development of this pipeline also believes in its transformative impact from a social, technological, and economic point of view.

Social Impact By enabling individuals to record various types of data remotely using everyday devices, the proposed pipeline facilitates non-invasive diagnostic procedures, eliminating the need for expensive and invasive tests. The accessibility of the proposed pipeline extends beyond traditional healthcare settings, allowing individuals in remote or underserved areas to access diagnostic services conveniently. Individuals can be monitored remotely, allowing healthcare providers to track their health status in real-time and intervene promptly if abnormalities are detected [36]. Early detection and intervention facilitated by the pipeline lead to improved prognoses for patients, as healthcare providers can initiate treatment at earlier stages of disease progression.

Technological Impact The proposed framework has significant technological implications. By leveraging the adaptable nature of the framework, models developed for one medical application can be readily deployed and tested in related contexts, accelerating the pace of medical research and innovation. As an example, the presented voice disorder detection model could be tested for neurodegenerative patients. Additionally, the framework enables the fine-tuning of models for specific cases, including those with limited data availability, such as rare diseases. Despite the scarcity of data in such contexts, the model can still generalize due to its original training on a larger and diverse dataset.

Economic Impact From an economic point of view, the proposed pipeline has low operating costs due to the utilization of personal devices for data collection and cloud-based analysis. Additionally, the framework facilitates the collection of low-cost, virtuous-cycle data, enabling continuous monitoring and feedback loops that enhance the accuracy and effectiveness of diagnostic and monitoring processes over time. Moreover, the framework lowers the burden of diagnosis and monitoring on public healthcare structures. This, in turn, enables healthcare professionals to focus on more complex and critical medical issues, ultimately improving the efficiency and resource allocation of the healthcare system.

References

[1] P. Rajpurkar, E. Chen, O. Banerjee, E. J. Topol, Ai in health and medicine, Nature medicine 28 (2022) 31–38.

[2] J.-R. Rueda, I. Sola, A. Pascual, M. S. Casacuberta, Non-invasive interventions for improving well-being and quality of life in patients with lung cancer, Cochrane Database of Systematic Reviews (2011).

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

[4] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, nature 521 (2015) 436–444.

[5] A. Baevski, Y. Zhou, A. Mohamed, M. Auli, wav2vec 2.0: A framework for self-supervised learning of speech representations, Advances in neural information processing systems 33 (2020) 12449–12460.

[6] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, A. Mohamed, Hubert: Self-supervised speech representation learning by masked prediction of hidden units, IEEE/ACM Transactions on Audio, Speech, and Language Processing 29 (2021) 3451–3460.

[7] H. Xiao, L. Li, Q. Liu, X. Zhu, Q. Zhang, Transformers in medical image segmentation: A review, Biomedical Signal Processing and Control 84 (2023) 104791.

[8] M. La Quatra, L. Vaiani, A. Koudounas, L. Cagliero, P. Garza, E. Baralis, How much attention should we pay to mosquitoes?, in: Proceedings of the 30th ACM International Conference on Multimedia, MM '22, Association for Computing Machinery, New York, NY, USA, 2022, p. 7135–7139. URL: https://doi.org/10.1145/3503161.3551594. doi:10.1145/3503161.3551594.

[9] A. Rajkomar, E. Oren, K. Chen, A. M. Dai, N. Hajaj, M. Hardt, P. J. Liu, X. Liu, J. Marcus, M. Sun, et al., Scalable and accurate deep learning with electronic health records, NPJ digital medicine 1 (2018) 18.

[10] K. Suzuki, Overview of deep learning in medical imaging, Radiological physics and technology 10 (2017) 257–273.

[11] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, et al., Highly accurate protein structure prediction with alphafold, Nature 596 (2021) 583–589.

[12] M. Lovino, V. Randazzo, G. Ciravegna, P. Barbiero, E. Ficarra, G. Cirrincione, A survey on data integration for multi-omics sample clustering, Neurocomputing 488 (2022) 494–508.

[13] L. Vaiani, A. Koudounas, M. La Quatra, L. Cagliero, P. Garza, E. Baralis, Transformer-based non-verbal emotion recognition: Exploring model portability across speakers' genders, in: Proceedings of the 3rd International on Multimodal Sentiment Analysis Workshop and Challenge, MuSe' 22, Association for Computing Machinery, New York, NY, USA, 2022, p. 89–94. URL: https://doi.org/10.1145/3551876.3554801. doi:10.1145/3551876.3554801.

[14] X. Liu, L. Xie, Y. Wang, J. Zou, J. Xiong, Z. Ying, A. V. Vasilakos, Privacy and security issues in deep learning: A survey, IEEE Access 9 (2020) 4566–4593.

[15] T. Madiega, Artificial intelligence act, European Parliament: European Parliamentary Research Service (2021).

[16] A. Vellido, The importance of interpretability and visualization in machine learning for applications in medicine and health care, Neural computing and applications 32 (2020).

[17] A. Koudounas, E. Pastor, G. Attanasio, V. Mazzia, M. Giollo, T. Gueudre, L. Cagliero, L. de Alfaro, E. Baralis, D. Amberti, Exploring subgroup performance in end-to-end speech models, in: ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2023, pp. 1–5. doi:10.1109/ICASSP49357.2023.10095284.

[18] A. Koudounas, E. Pastor, G. Attanasio, V. Mazzia, M. Giollo, T. Gueudre, E. Reale, L. Cagliero, S. Cumani, L. de Alfaro, E. Baralis, D. Amberti, Towards comprehensive subgroup performance analysis in speech models, IEEE/ACM Transactions on Audio, Speech, and Language Processing (2024). doi:10.1109/TASLP.2024.3363447.

[19] A. Koudounas, F. Giobergia, E. Baralis, Bad exoplanet! explaining degraded performance when reconstructing exoplanets atmospheric parameters, in: NeurIPS 2023 AI for Science Workshop, 2023. URL: https://openreview.net/forum?id=9Z4XZOhwiz.

[20] G. Vilone, L. Longo, Explainable artificial intelligence: a systematic review, arXiv preprint arXiv:2006.00093 (2020).

[21] E. Pastor, A. Koudounas, G. Attanasio, D. Hovy, E. Baralis, Explaining speech classification models via word-level audio segments and paralinguistic features, in: Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, 2024.

[22] G. Ciravegna, P. Barbiero, F. Giannini, M. Gori, P. Lió, M. Maggini, S. Melacci, Logic explained networks, Artificial Intelligence 314 (2023) 103822.

[23] P. Barbiero, G. Ciravegna, F. Giannini, M. Espinosa Zarlenga, L. C. Magister, A. Tonda, P. Lio, F. Precioso, M. Jamnik, G. Marra, Interpretable neural-symbolic concept reasoning, in: A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, J. Scarlett (Eds.), Proceedings of the 40th International Conference on Machine Learning, volume 202 of *Proceedings of Machine Learning Research*, PMLR, 2023, pp. 1801–1825.

[24] M. Espinosa Zarlenga, P. Barbiero, G. Ciravegna, G. Marra, F. Giannini, M. Diligenti, Z. Shams, F. Precioso, S. Melacci, A. Weller, et al., Concept embedding models: Beyond the accuracy-explainability trade-off, Advances in Neural Information Processing Systems 35 (2022) 21400–21413.

[25] E. Poeta, G. Ciravegna, E. Pastor, T. Cerquitelli, E. Baralis, Concept-based explainable artificial intelligence: A survey, arXiv preprint arXiv:2312.12936 (2023).

[26] N. Roy, R. M. Merrill, S. D. Gray, E. M. Smith, Voice disorders in the general population: prevalence, risk factors, and occupational impact, The Laryngoscope 115 (2005) 1988–1995.

[27] S. M. Cohen, Self-reported impact of dysphonia in a primary care population: An epidemiological study, The Laryngoscope 120 (2010) 2022–2032.

[28] N. Bhattacharyya, The prevalence of voice problems among adults in the united states, The Laryngoscope 124 (2014).

[29] N. Spantideas, E. Drosou, A. Karatsis, D. Assimakopoulos, Voice disorders in the general greek population and in patients with laryngopharyngeal reflux. prevalence and risk factors, Journal of Voice 29 (2015) 389–e27.

[30] E. Brunner, K. Eberhard, M. Gugatschka, Prevalence of benign vocal fold lesions: Long-term results from a single european institution, Journal of Voice (2023).

[31] I. Karabayir, S. M. Goldman, S. Pappu, O. Akbilgic, Gradient boosting for parkinson's disease diagnosis from voice recordings, BMC Medical Informatics and Decision Making 20 (2020).

[32] H. Vieira, N. Costa, T. Sousa, S. Reis, L. Coelho, Voice-based classification of amyotrophic lateral sclerosis: where are we and where are we going? a systematic review, Neurodegenerative Diseases 19 (2020) 163–170.

[33] R. Islam, E. Abdel-Raheem, M. Tarique, Voice pathology detection using convolutional neural networks with electroglottographic (egg) and speech signals, Computer Methods and Programs in Biomedicine Update 2 (2022) 100074.

[34] X. Xie, H. Cai, C. Li, Y. Wu, F. Ding, A voice disease detection method based on mfccs and shallow cnn, Journal of Voice (2023).

[35] M. Radfar, A. Mouchtaris, S. Kunzmann, End-to-End Neural Transformer Based Spoken Language Understanding, in: Proc. Interspeech 2020, 2020, pp. 866–870. doi:10.21437/Interspeech.2020-1963.

[36] D. Apiletti, E. Baralis, G. Bruno, T. Cerquitelli, Real-time analysis of physiological data to support medical applications, IEEE transactions on information technology in biomedicine 13 (2009) 313–321.