Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Evaluation of Voice Features Using Wav2vec2.0-Based Deep Learning as a Diagnostic Tool in Acromegaly

Seyma Aksoy
Ministry of Health, Hakkari State Hospital
Emre Saygili
emre.saygili@comu.edu.tr

Canakkale Onsekiz Mart University
Salih Akyel
University of Health Sciences, Basaksehir Cam and Sakura City Hospital
Mutlu Niyazoglu
University of Health Sciences, Basaksehir Cam and Sakura City Hospital
Esra Hatipoglu
University of Health Sciences, Basaksehir Cam and Sakura City Hospital

Research Article

Keywords: Acromegaly, Digital Biomarker, Machine Learning, Wav2vec2.0, Voice Analysis

Posted Date: July 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10395791/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

Acromegaly has an insidious onset and prolonged diagnostic delay. Vocal changes are a recognized clinical feature, the diagnostic value of deep learning architectures remains underexplored. We aimed to assess Wav2Vec2.0-based voice analysis as a non-invasive digital biomarker for acromegaly.

Methods

This single-center case-control study included 70 patients with acromegaly and 68 age- and sex-matched controls. Clinical status was defined by standard biochemical criteria, symptoms, comorbidities, and residual tumor. Voice samples were recorded with a standard mobile phone and resampled to 16 kHz. Wav2Vec2.0 was used to extract high-dimensional embedding vectors. These embeddings, were used to train and validate four different machine learning classifiers using an 80%/20% train–test split with five-fold cross-validation. Model calibration was assessed with the Brier score, and clinical utility with decision curve analysis (DCA).

Results

Demographics were similar between acromegaly (mean age 45.5 years; 57.1% female) and control (mean age 47.5 years; 61.8% female) groups. In five-fold cross-validation, Logistic Regression(LR) performed best, with an AUC of 0.86 (95% CI: 0.79–0.93), sensitivity 0.73 (0.57–0.90), and specificity 0.79 (0.72–0.87). On the test set, LR maintained high accuracy (AUC 0.84; specificity 0.92; sensitivity 0.71) and showed good calibration (Brier score 0.1753). DCA confirmed that LR provided the highest benefit across a clinically relevant range of threshold probabilities (~ 15–85%).

Conclusion

Voice analysis Wav2Vec2.0-based model demonstrates good diagnostic accuracy for acromegaly. These preliminary findings suggest that voice may serve as a promising non-invasive digital biomarker, pending validation in larger and more diverse cohorts.

Introduction

Acromegaly is a rare endocrine disorder typically characterized by the hypersecretion of GH and IGF-1, most commonly due to a pituitary adenoma. Due to the insidious onset and slow progression of clinical symptoms, achieving a timely diagnosis remains a significant challenge [1, 2]. Research indicates that the diagnostic delay in acromegaly can span from 2 to 25 years [3]. For instance, the ACRO-POLIS study, involving 472 patients, reported a substantial mean interval of 14.2 (± 11.3) years between the initial onset

of symptoms and diagnosis [4]. Such prolonged diagnostic delays are critically associated with increased mortality rates and a significant decline in patient quality of life [5]. Consequently, extensive research has focused on developing advanced diagnostic tools to facilitate both the early detection of acromegaly and the timely identification of disease recurrence [6].

Acromegaly is a rare multisystemic disorder that necessitates lifelong clinical management due to its association with extensive comorbidities. These complications encompass a broad spectrum of cardiovascular, metabolic, and skeletal manifestations, as well as orofacial changes and soft tissue abnormalities [7]. Specifically, growth hormone (GH) promotes the proliferation of osteoblastic lineage cells and modulates the differentiation of mesenchymal progenitor cells, thereby stimulating both osteoblastogenesis and chondrogenesis [8]. When these processes involve the craniofacial skeleton, they result in the characteristic facial appearance pathognomonic of acromegaly.

Voice alterations in acromegaly result from complex orofacial and laryngeal structural modifications secondary to prolonged exposure to elevated GH levels. Cortical bone thickening involving the cranium significantly alters the resonance properties of the vocal tract [9]. Furthermore, hypertrophy of the paranasal sinuses -particularly the frontal sinuses- combined with laryngeal tissue expansion, typically leads to a deeper, more resonant vocal quality [10]. Characteristically, patients exhibit a lower vocal pitch and increased hoarseness, which are attributed to the thickening and increased laxity of the vocal folds [11]. While soft tissue hypertrophy may be partially reversible following the normalization of GH levels [12–14], the relationship between hormonal activity and certain comorbidities remains complex.

Previous studies utilizing digital acoustic voice analysis in patients with acromegaly have identified significant abnormalities in vocal components. These alterations are often clinically perceived as dysphonia and may manifest as persistent hoarseness, breathiness, or a characteristic deepening of the vocal pitch [15–17].

Digital voice data, when analyzed through a combination of artificial intelligence (AI) and machine learning (ML), holds significant potential as a biomarker for monitoring disease progression across various patient cohorts [18]. Despite the promising results of deep learning methodologies, several challenges persist, including the scarcity of large-scale, high-quality datasets and the subjective, time-consuming nature of feature engineering based on domain expertise [19]. To date, research in this field has primarily focused on Alzheimer’s disease, Parkinson’s disease, and depression [19–22].

While numerous studies have applied machine learning to facilitate the diagnosis of acromegaly through facial photography [23–26], research integrating digital voice analysis with ML remains extremely limited. To date, only a single study, by Vouzouneraki et al., has explored this intersection, using a classical machine learning approach [27].

Voice alterations are a clinically significant and frequent manifestation of acromegaly that substantially impair patient quality of life. We therefore hypothesized that acromegaly-specific vocal signatures could serve as a reliable, non-invasive digital biomarker. Capturing such signatures, however, requires methods capable of extracting fine-grained acoustic detail.

Recent advances in speech analysis offer such methods. Self-supervised architectures such as Wav2Vec 2.0 (Wave-to-Vector), Whisper, WavLM (Waveform Large-scale Model), and HuBERT (Hidden-unit BERT) are pre-trained on vast quantities of unlabeled audio data [28, 29]. Among these, Wav2Vec 2.0 focuses on extracting acoustic properties -such as raw waveforms and spectral features- rather than parsing linguistic or semantic content, and its reduced reliance on large labeled datasets makes it well suited for acoustic feature extraction within multimodal frameworks [22].

Building on these capabilities, the present study applied transfer learning -a technique that leverages knowledge from a pre-trained model to improve learning in a target domain [30]- to identify disease-specific acoustic characteristics using the Wav2Vec 2.0 architecture. To our knowledge, this is the first application of this approach in acromegaly. We aimed to evaluate its diagnostic utility and to establish voice analysis as a non-invasive screening tool for clinical practice.

Materials and Methods

Study Population and Ethical Approval

This study included a total of 138 participants, comprising 70 patients diagnosed with acromegaly who presented to the endocrinology clinic between January 2024 and June 2025 and 68 age- and sex-matched control subjects with non-functioning pituitary adenomas. The acromegaly cohort included patients in both biochemically active and remission states. All participants underwent comprehensive laryngoscopic examination by an ear, nose, and throat specialist for the assessment of vocal fold thickness. In addition, the SAGIT score, which incorporates signs and symptoms, associated comorbidities, GH levels, IGF-1 levels, and tumor characteristics, was calculated for each patient in the acromegaly cohort to evaluate disease activity and severity. The study protocol was reviewed and approved by the Local Institutional Ethics Committee (Approval No: 711, Date: December 27, 2023) and was conducted in accordance with the ethical principles of the Declaration of Helsinki. Before enrollment, written informed consent was obtained from all participants after a detailed explanation of the study objectives and procedures.

Audio Data Collection and Preprocessing

Voice data collection was conducted in a controlled, quiet examination room. Each participant was instructed to repeat the Turkish words ‘zaman’ and ‘simit’ three times [31]. To ensure consistency, a fixed mobile phone microphone position at a standard distance of 15 cm was utilized. Each participant provided exactly one continuous audio recording containing the concatenated repetitions, resulting in a total of 138 unique audio clips (70 acromegaly, 68 control) and a strict 1:1 participant-to-clip ratio. Audio preprocessing was implemented in Python using the librosa library (version 0.10.1) [32]. The recordings were loaded at their native 16 kHz sampling rate in mono channel format. To normalize length and limit computational load, the audio waveforms were truncated to a maximum duration of 15 seconds. No silence trimming or manual amplitude normalization was applied to the raw waveforms before feature extraction.

Classical Acoustic Feature Extraction

Acoustic features were extracted from preprocessed voice recordings using the praat-parselmouth library (version 0.4.7) [33], a Python interface for Praat [34]. Eleven acoustic parameters were analyzed: F0 descriptors, local jitter, local shimmer, Harmonics-to-Noise Ratio (HNR), and the first three formant frequencies (F1–F3). F0 was estimated by autocorrelation-based pitch detection, HNR by cross-correlation-based harmonicity analysis, and formants by the Burg linear predictive coding algorithm. Variables that could not be reliably estimated were treated as missing and excluded from the corresponding analyses without imputation.

Deep Feature Extraction via Wav2Vec 2.0

In addition to conventional acoustic parameters, deep speech features were extracted using the Wav2Vec 2.0 architecture [35]. The multilingual facebook/wav2vec2-large-xlsr-53 model (36) was used as a fixed feature extractor without further training. Feature extraction was performed locally using PyTorch (version 2.10.0) [37] and the Hugging Face Transformers library (version 4.44.2) [38]. Each 15-second mono voice recording was processed by the model, and the resulting time-dependent hidden representations were averaged over time to obtain a single 1024-dimensional deep acoustic feature vector for each participant.

Machine Learning Classification Pipeline

The development and reporting of the diagnostic prediction models adhered to the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis + Artificial Intelligence (TRIPOD + AI) statement [39]. A machine learning classification framework was developed using scikit-learn (version 1.7.2) [40] to assess the diagnostic utility of the deep embeddings. Four algorithms were evaluated: Logistic Regression (LR) [41], Support Vector Machine (SVM) with a radial basis function kernel (42), Random Forest (RF) [43], and Extreme Gradient Boosting (XGBoost; version 2.1.4) [44]. The dataset of 138 participants was split into a stratified training set (80%, N = 110; 56 acromegaly, 54 controls) and an internal hold-out test set (20%, N = 28; 14 acromegaly, 14 controls). Feature scaling with StandardScaler was incorporated into a scikit-learn Pipeline to prevent data leakage. Hyperparameter optimization was conducted on the training set using randomized search combined with stratified 5-fold cross-validation. The final model was evaluated on the internal hold-out test set, and probability calibration of the best-performing model, Logistic Regression, was assessed using the Brier score [45].

Statistical Analysis

Continuous demographic and acoustic parameters were summarized as median values with interquartile ranges (Q1–Q3) as their distribution deviated from normality. Group comparisons between the acromegaly and control cohorts were conducted using the non-parametric Mann–Whitney U test. Classification performance was evaluated using Accuracy, AUC, Sensitivity (Recall), Specificity, Precision, F1-score, Cohen’s Kappa, and Matthews Correlation Coefficient. All statistical analyses were performed using

Python software (version 3.10.12) utilizing Pandas (version 2.3.3) for data management and SciPy (version 1.15.3) for statistical testing. All statistical tests were two-sided, with statistical significance defined as $p < 0.05$.

Results

Patient Characteristics

The final cohort included 138 participants: 70 patients with acromegaly and 68 controls with non-functioning pituitary adenomas. The acromegaly group comprised patients with either biochemically controlled or biochemically active disease. Median age was 45.5 years in the acromegaly group and 47.5 years in the control group. Females accounted for 40 patients (57.1%) in the acromegaly group and 42 (61.8%) in the control group, while males accounted for 30 (42.9%) and 26 (38.2%), respectively. Age and sex distribution did not differ significantly between groups ($p = 0.44$ and $p = 0.58$, respectively). Vocal cord thickness did not differ between the control and acromegaly groups (5.0 mm [Q1–Q3: 4.4–5.0] vs 5.0 mm [Q1–Q3: 5.0–5.0], respectively; $p = 0.14$) (Table 1).

Table 1
Comparison of demographic and laryngeal parameters between the acromegaly and control groups.

Characteristic	Control (n = 68)	Acromegaly (n = 70)	p
Age (years)	47.5 (40.4–58.0)	45.5 (39.8–53.1)	0.44
Females, n (%)	42 (61.8%)	40 (57.1%)	0.58
Vocal cord thickness (mm)	5.0 (4.4–5.0)	5.0 (5.0–5.0)	0.14
Values are expressed as median (Q1–Q3) unless otherwise stated.			

Within the acromegaly cohort, patients represented both newly diagnosed cases and individuals under long-term follow-up. Disease duration ranged from 1 month to 37 years, with a median of 62 months. 42 patients (60%) were classified as biochemically controlled, whereas 28 (40%) had biochemically active disease. Sex distribution was comparable between disease-status subgroups; females represented 59.5% of the biochemically controlled group and 53.6% of the active disease group ($p = 0.62$).

Biochemical and clinical parameters differed according to disease status. Median IGF-1 was 1.0 × upper limit of normal (ULN) (Q1–Q3: 1.0–1.0) in the biochemically controlled group and 1.8 × ULN (Q1–Q3: 1.4–2.8) in the active disease group ($p < 0.01$). The prevalence of hypertension was similar between groups ($p = 0.81$). Surgical history differed significantly, with 40 patients (95.2%) in the biochemically controlled group and 20 (71.4%) in the active disease group having undergone surgery ($p = 0.01$). Median tumour size was greater in the active disease group (11.5 mm [Q1–Q3: 6.3–20.7]) than in the biochemically controlled group (0.0 mm [Q1–Q3: 0.0–0.0]; $p < 0.01$). SAGIT scores were also higher among patients with active disease (median 8.0; Q1–Q3: 6.4–11.6) than among those with biochemically controlled disease (median

4.0; Q1–Q3: 2.9–4.0; p < 0.01). Radiotherapy exposure did not differ significantly between groups (p = 0.33) (Table 2).

Table 2
Comparison of demographic, biochemical and laryngeal parameters according to biochemical disease status in acromegaly patients.

Characteristic	Biochemically controlled (n = 42)	Biochemically active (n = 28)	p
Females, n (%)	25 (59.5%)	15 (53.6%)	0.62
Males, n (%)	17 (40.5%)	13 (46.4%)	0.62
Vocal cord thickness (mm)	5.0 (5.0–5.0)	5.0 (4.4–5.0)	0.94
IGF-1 (× ULN)	1.0 (1.0–1.0)	1.8 (1.4–2.8)	< 0.01
Hypertension, n (%)	12 (28.6%)	7 (25.9%)	0.81
Surgery, n (%)	40 (95.2%)	20 (71.4%)	0.01
Radiotherapy, n (%)	3 (7.1%)	4 (14.3%)	0.33
Tumour size (mm)	0.0 (0.0–0.0)	11.5 (6.3–20.7)	< 0.01
SAGIT score	4.0 (2.9–4.0)	8.0 (6.4–11.6)	< 0.01

IGF-1: insulin-like growth factor 1; ULN: upper limit of normal. Values are expressed as median (Q1–Q3) or n (%). Hypertension percentages are based on the patients with available comorbidity data.

Acoustic parameters

Baseline acoustic analysis demonstrated lower central F_0 measures in the acromegaly group compared with controls; however, the differences in F_0 mean and F_0 median did not reach statistical significance (p = 0.079 and p = 0.097, respectively). F_0 minimum was comparable between groups (p = 0.398), whereas F_0 maximum was significantly higher in the control group than in the acromegaly group (p = 0.032).

Among vocal perturbation parameters, median jitter and shimmer were numerically lower in the acromegaly group, although neither difference reached statistical significance (p = 0.162 and p = 0.073, respectively). HNR differed significantly between groups, with a higher median value in the acromegaly group than in controls (p = 0.007).

Taken together, conventional acoustic parameters showed a pattern characterized by lower F_0-related measures and higher HNR in the acromegaly group, although only F_0 maximum and HNR reached statistical significance (Table 3).

Table 3
Median (Q1–Q3) values of baseline voice parameters in acromegaly and control groups

Acoustic parameter	Acromegaly (n = 70)	Control (n = 68)	p
F_0 mean (Hz)	160.64 (120.95–189.66)	173.39 (128.04–197.76)	0.079
F_0 median (Hz)	166.21 (115.37–189.87)	173.53 (128.13–199.54)	0.097
F_0 minimum (Hz)	105.98 (86.01–158.87)	98.55 (83.56–140.82)	0.398
F_0 maximum (Hz)	207.87 (141.61–236.16)	225.10 (159.88–266.02)	0.032 *
Jitter (local)	0.010 (0.008–0.015)	0.012 (0.009–0.016)	0.162
Shimmer (local)	0.058 (0.047–0.085)	0.072 (0.055–0.085)	0.073
HNR (dB)	14.71 (12.90–16.72)	13.72 (12.08–14.73)	0.007 *

*F_0: fundamental frequency; HNR: harmonics-to-noise ratio. Values are median (Q1–Q3). Group comparisons by Mann–Whitney U test. * p < 0.05.*

Machine Learning-Based Voice Analysis

To assess whether voice recordings could discriminate patients with acromegaly from controls, recordings from all participants were converted into high-dimensional Wav2Vec 2.0 embedding vectors and used as input features for four classifiers: Logistic Regression (LR), Support Vector Machine (SVM), Extreme Gradient Boosting (XGBoost) and Random Forest (RF). Model development was performed using five-fold cross-validation within the training cohort, followed by independent evaluation in the hold-out test cohort.

5-Fold Cross-Validation Results

During five-fold cross-validation, LR demonstrated the most consistent overall performance. It achieved the highest mean AUC (0.86; 95% CI: 0.79–0.93), accuracy (0.76; 95% CI: 0.68–0.85), sensitivity (0.73; 95% CI: 0.57–0.90), specificity (0.79; 95% CI: 0.72–0.87), precision (0.79; 95% CI: 0.73–0.85), F1 score (0.75; 95% CI: 0.66–0.85), Cohen's kappa (0.53; 95% CI: 0.36–0.70), and MCC (0.54; 95% CI: 0.37–0.71) among the evaluated models.

SVM showed the second-highest discrimination, with an AUC of 0.80 (95% CI: 0.70–0.89) and an MCC of 0.43 (95% CI: 0.27–0.59). XGBoost yielded an AUC of 0.77 (95% CI: 0.66–0.89), sensitivity of 0.67 (95% CI: 0.47–0.86), and MCC of 0.37 (95% CI: 0.19–0.56). RF showed the lowest cross-validation performance, with an AUC of 0.71 (95% CI: 0.55–0.88) and an MCC of 0.32 (95% CI: 0.10–0.55) (Table 4).

Table 4
Five-fold cross-validation performance of the machine learning models in the training cohort (n = 110).

Model	Accuracy	AUC	Sensitivity	Specificity	Precision	F1	Kappa	MCC
LR	0.76 (0.68–0.85)	0.86 (0.79–0.93)	0.73 (0.57–0.90)	0.79 (0.72–0.87)	0.79 (0.73–0.85)	0.75 (0.66–0.85)	0.53 (0.36–0.70)	0.54 (0.37–0.71)
SVM	0.71 (0.63–0.79)	0.80 (0.70–0.89)	0.72 (0.58–0.85)	0.70 (0.55–0.86)	0.73 (0.63–0.83)	0.71 (0.63–0.79)	0.42 (0.26–0.58)	0.43 (0.27–0.59)
XGBoost	0.67 (0.59–0.76)	0.77 (0.66–0.89)	0.67 (0.47–0.86)	0.69 (0.53–0.84)	0.70 (0.61–0.80)	0.66 (0.56–0.76)	0.35 (0.18–0.52)	0.37 (0.19–0.56)
RF	0.65 (0.55–0.76)	0.71 (0.55–0.88)	0.65 (0.45–0.85)	0.67 (0.61–0.72)	0.66 (0.59–0.73)	0.64 (0.52–0.77)	0.31 (0.10–0.52)	0.32 (0.10–0.55)

LR: logistic regression; SVM: support vector machine; XGBoost: extreme gradient boosting; RF: random forest; AUC: area under the ROC curve; F1: F1 score; Kappa: Cohen's kappa; MCC: Matthews correlation coefficient. Results are reported as mean values with 95% confidence intervals.

Test Set Results

In the independent test cohort, LR achieved the highest AUC (0.847), specificity (0.929), and precision (positive predictive value, 0.909). Its accuracy, sensitivity, F1 score, Cohen's kappa, and MCC were 0.821, 0.714, 0.800, 0.643, and 0.658, respectively. XGBoost achieved the highest accuracy (0.857), sensitivity (0.857), F1 score (0.857), Cohen's kappa (0.714), and MCC (0.714), with an AUC of 0.811 and specificity of 0.857. SVM showed intermediate performance (AUC 0.827; accuracy 0.786), whereas RF had the lowest test-set performance (AUC 0.724; MCC 0.144) (Table 5).

Table 5
Performance of the machine learning models in the independent test cohort (n = 28).

Model	Accuracy	AUC	Sensitivity	Specificity	Precision	F1	Kappa	MCC
LR	0.821	0.847	0.714	0.929	0.909	0.800	0.643	0.658
SVM	0.786	0.827	0.714	0.857	0.833	0.769	0.571	0.577
XGBoost	0.857	0.811	0.857	0.857	0.857	0.857	0.714	0.714
RF	0.571	0.724	0.500	0.643	0.583	0.538	0.143	0.144

LR: logistic regression; SVM: support vector machine; XGBoost: extreme gradient boosting; RF: random forest; AUC: area under the ROC curve; F1: F1 score; Kappa: Cohen's kappa; MCC: Matthews correlation coefficient. Values are point estimates on the hold-out test set.

Across cross-validation and hold-out testing, LR provided the most stable discriminatory performance, with an AUC of 0.86 in cross-validation and 0.84 in the test cohort, together with high specificity and

positive predictive value in the independent test set. On this basis, LR was selected as the primary classifier for subsequent calibration and decision-curve analyses.

Probability Calibration of the Logistic Regression Classifier

Probability calibration was subsequently evaluated for the LR classifier in the test cohort. The model yielded a Brier score of 0.1753, indicating adequate calibration of predicted probabilities.

Decision Curve Analysis (DCA)

DCA was performed to quantify the clinical net benefit of each classifier across threshold probabilities from 1% to 99%. The LR model showed the highest net benefit relative to SVM, XGBoost, RF, and the default treat-all and treat-none strategies. A positive net benefit for LR was observed across a broad range of threshold probabilities, approximately 15% to 85%.

Discussion

The primary finding of this study is that LR model consistently superiorly outperformed the other three machine learning models evaluated. In the LR model high diagnostic accuracy maintained across both training and independent testing phases (AUC 0.86 vs. 0.84) confirms the model's robustness and the absence of overfitting. Crucially, this performance was validated on data entirely external to the training phase, which the model had not previously encountered. The consistency between the cross-validation results and the test set performance confirms the absence of overfitting, ensuring that the model effectively to unseen patient data.

A key technical contribution of our work is the application of transfer learning through the Wav2Vec2.0 architecture. By leveraging pre-trained representations and fine-tuning them on our specific dataset, we successfully developed a high-performing classification model despite the inherent constraints of a relatively small sample size. This approach allowed the model to utilize features learned from vast amounts of diverse speech data and repurpose them for the specialized task of identifying acromegalic vocal changes. Consequently, our findings suggest that deep learning-based embeddings, even with limited clinical data, can provide a diagnostic biomarker.

While Wav2Vec 2.0 is a complex deep learning architecture, we used it solely as a feature extractor to generate fixed-length embedding vectors, not as an end-to-end classifier. The final diagnostic decision was made by LR, which -unlike deep neural networks that are often criticized as "black boxes" with opaque internal decision processes- is a transparent, white-box model. LR assigns an explicit coefficient to each input feature, expressed as an odds ratio, allowing the contribution of every variable to the model's output to be quantified directly. For the hand-crafted acoustic features (e.g., jitter, shimmer, F_0, HNR), these coefficients carry direct physiological meaning and thus provide clinically interpretable insight. For the Wav2Vec 2.0 embeddings, the coefficients quantify the contribution of individual latent dimensions, which do not have a direct clinical correlate. We therefore interpret feature contributions on the log-odds scale and rely primarily on the acoustic-feature coefficients for clinical interpretation [46].

Our findings demonstrate a diagnostic performance that is comparable to previous machine learning applications in acromegaly. In a notable study by Vouzouneraki et al., the authors utilized an ensemble of Support Vector Machines (SVM), Random Forest (RF), and k-Nearest Neighbors (KNN) models, trained on 3,274 acoustic features from a larger cohort of 290 participants (n = 151 acromegaly, n = 139 control). Their model ensemble yielded a sensitivity of 0.71, a specificity of 0.77, and an AUC of 0.84 [27]. In comparison, our study achieved a similar AUC (0.84) and sensitivity (0.71), while demonstrating a notably higher specificity (0.92), despite having a smaller sample size (n = 70 acromegaly, n = 68 control). This efficiency is largely attributed to our methodology: while traditional acoustic parameters were used for group comparisons, the machine learning component utilized the Wav2Vec2.0 model for feature extraction. This approach facilitated the generation of 1024-dimensional embedding vectors, providing a more comprehensive representation of vocal characteristics than traditional features.

Early researches have established the premise that digital voice analysis, integrated with machine learning, can serve as a diagnostic tool for cardiovascular, neurological, respiratory, and mental diseases [47]. A review of existing literature on machine learning-based speech analysis indicates that the Wav2Vec2.0 technique is frequently employed for vocal assessment [22, 48, 49]. This architecture enables machine learning models to learn robust speech representations from unlabeled data by predicting masked or missing segments of the input sequence [48]. By utilizing multiple encoding layers, Wav2Vec2.0 extracts meaningful features from raw audio signals, facilitating more efficient data processing for downstream clinical classification tasks [22].

An examination of literature utilizing the Wav2Vec2.0 framework alongside machine learning reveals varying performance outcomes across different deep learning models. For instance, in a study by Dudek et al. regarding Parkinson's disease, the most favorable measurement values were achieved using the XGBoost model [28]. Similarly, research by Kiyoshige et al. demonstrated high predictive performance (AUC: 0.89) when using XGBoost and Deep Neural Network (DNN) models to identify cognitive impairment through vocal biomarkers [50]. In our study, the highest average AUC value was obtained with the LR model, reaching 0.86 (95% CI: 0.79–0.93).

Traditional acoustic analysis was also carried out within our cohort. When comparing these findings to the existing literature, certain parallels and distinctions emerge. Specifically, while Bogazzi et al. reported a significant decrease in the fundamental frequency (F_0) among acromegaly patients, the maximum F_0 in our study was significantly higher in the control group compared to the acromegaly group ($p = 0.032$). Our findings regarding the HNR were consistent with previous research. The median HNR was significantly higher in the acromegaly group in our study ($p = 0.007$). These changes in the harmonic structure likely reflect the structural modifications in supraglottic resonance space characteristic of the disease [17].

In patients with acromegaly, chronic hormonal elevations lead to characteristic dental and craniofacial structural alterations [51]. The most prominent changes typically manifest in the facial bones, specifically through the bossing of the frontal bone and a measurable increase in the anterior vertical facial height [9, 52]. Previous evaluations have identified the mandible as the site of the most significant abnormality, where the ascending ramus exhibits a more pronounced growth pattern compared to the mandibular body

[53]. Although chronic GH and IGF-1 excesses are typically associated with soft tissue hypertrophy, our study did not find a statistically significant difference in vocal fold thickness between the groups. Our findings suggest that the changes in voice parameters -such as the significant increase in HNR- might be driven more by changes in the resonance chambers (sinuses, oral cavity, and pharyngeal space) rather than the physical thickness of the vocal folds themselves.

Digital biomarkers are experiencing rapid advancement within the field of digital medicine by using smartphones or wearable digital systems [54–58]. Speech, in particular, provides a wealth of information regarding cognitive function and is frequently affected by various psychiatric and neurodegenerative conditions [54, 59]. While previous research in acromegaly has explored the development of diagnostic algorithms using facial photographs, only one prior study has utilized voice analysis for this purpose. Although digital biomarkers may not immediately shift the overall medical paradigm, these novel methods of measuring health status offer observations and perspectives into disease states that were previously unavailable [60].

A key methodological strength of our work is its pioneering approach to acromegaly screening. To our knowledge, this is the second study to investigate digital voice parameters as a potential biomarker for the machine learning-based diagnosis of acromegaly, and the first to apply the Wav2Vec 2.0 architecture to voice analysis in this patient population.

Our study has limitations that should be acknowledged. First, specific clinical information regarding the participants' active smoking status and recent asthma treatments was not available, which are factors that may influence vocal quality. Second, the study cohort included acromegaly patients at various clinical stages -ranging from recent diagnosis to biochemical remission- a factor that potentially affects the degree of vocal alteration. And also due to the relatively small sample size within each clinical subgroup, we were unable to conduct a comparative analysis of voice parameters between patients in biochemical remission and those with active disease. A further limitation is that while Wav2vec2.0 provides high-dimensional feature extraction, its 'black-box' nature lacks the transparent clinical interpretability and physiological counterparts found in traditional acoustic measurements. Consequently, the potential influence of current hormonal activity on specific acoustic features remains to be elucidated in larger, stratified cohorts. Furthermore, as acromegaly is a rare disease, the sample size was relatively small, which inherently limits the amount of data available for training machine learning models for diagnostic purposes. By utilizing the Wav2vec 2.0 model and leveraging the power of transfer learning, we were able to develop a machine learning model that achieved a high AUC rate despite the smaller patient cohort.

Conclusion

Machine learning-based analysis of mobile phone recorded voice samples may offer a promising, non-invasive approach for the detection of acromegaly. Using Wav2Vec 2.0-derived speech embeddings, LR showed the most consistent diagnostic performance, with favorable discrimination in both cross-validation and independent test analyses, along with high specificity and positive predictive value. These preliminary findings support the potential utility of voice-derived digital biomarkers as a complementary

screening tool for acromegaly. Further validation in larger, clinically diverse, and multilingual cohorts is required to confirm the robustness, generalizability, and clinical applicability of this approach.

Declarations

Conflict of interest:

The authors declare that they have no conflict of interest.

AI Disclosure

The authors would like to acknowledge the use of Gemini version 3.0 (Google LLC, Mountain View, CA, USA) for English language editing and proofreading to enhance the clarity of the manuscript. The final text was thoroughly reviewed and approved by all authors, who maintain full accountability for the integrity of the work.

Funding:

This research did not receive any specific grant from any funding agency in the public, commercial or not-for-profit sector.

Author Contribution

S. [Aksoy] collected the data and drafted the manuscript. E.S. performed the data analyses and drafted the manuscript. S. [Akyel] contributed to data collection. All authors reviewed and approved the final manuscript.

References

1. Danilowicz K, Day PF, Manavela MP, Herrera CJ, Deheza ML, Isaac G et al (2016) Implementing a screening program for acromegaly in Latin America: necessity versus feasibility. Pituitary 19(4):370–374. 10.1007/s11102-016-0714-5
2. Rajasoorya C, Holdaway IM, Wrightson P, Scott DJ, Ibbertson HK (1994) Determinants of clinical outcome and survival in acromegaly. Clin Endocrinol (Oxf) 41(1):95–102. 10.1111/j.1365-2265.1994.tb03789.x PubMed PMID: 8050136
3. Petrossians P, Daly AF, Natchev E, Maione L, Blijdorp K, Sahnoun-Fathallah M et al (2017) Acromegaly at diagnosis in 3173 patients from the Liège Acromegaly Survey (LAS) Database. Endocr Relat Cancer 24(10):505–518. 10.1530/ERC-17-0253 PubMed PMID: 28733467; PubMed Central PMCID: PMC5574208

4. Caron P, Brue T, Raverot G, Tabarin A, Cailleux A, Delemer B et al (2019) Signs and symptoms of acromegaly at diagnosis: the physician's and the patient's perspectives in the ACRO-POLIS study. Endocrine 63(1):120–129. 10.1007/s12020-018-1764-4
5. Chiloiro S, Giampietro A, Gagliardi I, Bondanelli M, Veleno M, Ambrosio MR et al (2022) Impact of the diagnostic delay of acromegaly on bone health: data from a real life and long term follow-up experience. Pituitary 25(6):831–841. 10.1007/s11102-022-01266-4
6. A W SC (2022) Questionnaire and tools: clinical powerful instrument in acromegaly diagnosis and management. J Endocrinol Invest 45(10). 10.1007/s40618-022-01782-x
7. Ben-Shlomo A, Sheppard MC, Stephens JM, Pulgar S, Melmed S (2011) Clinical, quality of life, and economic value of acromegaly disease control. Pituitary 14(3):284–294. 10.1007/s11102-011-0310-7
8. Claessen KMJA, Mazziotti G, Biermasz NR, Giustina A (2016) Bone and Joint Disorders in Acromegaly. Neuroendocrinology 103(1):86–95. 10.1159/000375450
9. Dogruel F, Canger Em G, Zb, Asantogrol F, Şahin A, Bayram F (2022) The evaluation of changes in maxillofacial bones using cone beam tomography in acromegaly. Med Oral Patol Oral Cir Bucal e357–e365. 10.4317/medoral.25280
10. De Stefani A, Dassie F, Wennberg A, Preo G, Muneratto A, Fabris R et al (2022) Oral Manifestations and Maxillo-Facial Features in the Acromegalic Patient: A Literature Review. J Clin Med 11(4):1092. 10.3390/jcm11041092
11. Edge WG, Whitwam JG (1981) Chondro-calcinosis and difficult intubation in acromegaly. Anaesthesia 36(7):677–680. 10.1111/j.1365-2044.1981.tb08779.x PubMed PMID: 7270844
12. Fatti LM, Scacchi M, Pincelli AI, Lavezzi E, Cavagnini F (2001) Prevalence and pathogenesis of sleep apnea and lung disease in acromegaly. Pituitary 4(4):259–262. 10.1023/a:1020702631793 PubMed PMID: 12501976
13. Kreitschmann-Andermahr I, Kohlmann J, Kleist B, Hirschfelder U, Buslei R, Buchfelder M et al (2018) Oro-dental pathologies in acromegaly. Endocrine 60(2):323–328. 10.1007/s12020-018-1571-y
14. Wolters TLC, Roerink SHPP, Drenthen LCA, Wagenmakers MAEM, Van Den Broek GB, Rutten KIM et al (2021) Voice Characteristics in Patients with Acromegaly during Treatment. J Voice 35(6):932. e13-932.e27
15. Aydin K, Turkyilmaz D, Ozturk B, Dagdelen S, Ozgen B, Unal F et al (2013) Voice characteristics of acromegaly. Eur Arch Otorhinolaryngol 270(4):1391–1396. 10.1007/s00405-013-2369-4
16. Weinberg B, Dexter R, Horii Y (1975) Selected speech and fundamental frequency characteristics of patients with acromegaly. J Speech Hear Disord 40(2):253–259. 10.1044/jshd.4002.253
17. Bogazzi F, Nacci A, Campomori A, La Vela R, Rossi G, Lombardi M et al (2010) Analysis of voice in patients with untreated active acromegaly. J Endocrinol Invest 33(3):178–185. 10.1007/BF03346578
18. Sara JDS, Orbelo D, Maor E, Lerman LO, Lerman A (2023) Guess What We Can Hear—Novel Voice Biomarkers for the Remote Detection of Disease. Mayo Clin Proc. ;98(9):1353–75. 10.1016/j.mayocp.2023.03.007

19. Huang X, Wang F, Gao Y, Liao Y, Zhang W, Zhang L et al (2024) Depression recognition using voice-based pre-training model. Sci Rep 14(1):12734. 10.1038/s41598-024-63556-0
20. Malekroodi HS, Madusanka N, Lee B, il YM (2025) Multi-Channel Spectro-Temporal Representations for Speech-Based Parkinson's Disease Detection. J Imaging 11(10):341. 10.3390/jimaging11100341
21. Klempir O, Mullerova JG, Krupicka R (2026) Statistical, multi-scale and attention-based layer pooling of Wav2Vec-2 speech embeddings for Parkinson's disease detection. Comput Biol Med 200:111368. 10.1016/j.compbiomed.2025.111368
22. Yang X, Hong K, Zhang D, Wang K (2024) Early diagnosis of Alzheimer's Disease based on multi-attention mechanism. Fati SM, editor. PLOS ONE. ;19(9):e0310966. 10.1371/journal.pone.0310966
23. Meng T, Guo X, Lian W, Deng K, Gao L, Wang Z et al (2020) Identifying Facial Features and Predicting Patients of Acromegaly Using Three-Dimensional Imaging Techniques and Machine Learning. Front Endocrinol 11:492. 10.3389/fendo.2020.00492
24. Kong X, Gong S, Su L, Howard N, Kong Y (2018) Automatic Detection of Acromegaly From Facial Photographs Using Machine Learning Methods. EBioMedicine 27:94–102. 10.1016/j.ebiom.2017.12.015
25. Kizilgul M, Karakis R, Dogan N, Bostan H, Yapici MM, Gul U et al (2023) Real-time detection of acromegaly from facial images with artificial intelligence. Eur J Endocrinol 188(1):lvad005. 10.1093/ejendo/lvad005
26. C YKXK, C H, L L (2020) Constructing an automatic diagnosis and severity-classification model for acromegaly using facial photographs by deep learning. J Hematol OncolJ Hematol Oncol 13(1). 10.1186/s13045-020-00925-y
27. Vouzouneraki K, Nylén F, Holmberg J et al (2025) Digital Voice Analysis as a Biomarker of Acromegaly. J Clin Endocrinol Metab 110(4):983–990. 10.1210/clinem/dgae689
28. Dudek M, Hemmerling D, Kaczmarska M, Stepien J, Daniol M, Wodzinski M et al (2025) Analysis of Voice, Speech, and Language Biomarkers of Parkinson's Disease Collected in a Mixed Reality Setting. Sensors 25(8):2405. 10.3390/s25082405
29. Kim D, Yi B, Won Y (2024) Speech Emotion Recognition in People at High Risk of Dementia. Dement Neurocognitive Disord 23(3):146. 10.12779/dnd.2024.23.3.146
30. Koval A, Mansouri SS, Kanellakis C (2023) Chap. 10 - Machine learning for ARWs, Aerial Robotic Workers. Butterworth-Heinemann, pp 159–174. doi.org/10.1016/B978-0-12-814909-6.00016-0
31. Yilmaz A, Kalebasi M, Erdinc B, Samoylenko Y, Inan D, Sisman B, Ozanalp C, Demirtas I, Yildiz U, Acar G, Ozdogan C, Karadamar N, Varol R, Cinier G, Tekkesin AI (2024) Machine learning based speech classification of patients with heart failure worsening, a pilot study [Abstract]. Eur J Heart Fail 26(Suppl 2):422–423
32. McFee B, Raffel C, Liang D, Ellis D, Mcvicar M, Battenberg E et al (2015) librosa: Audio and Music Signal Analysis in Python. 18. 10.25080/Majora-7b98e3ed-003
33. Jadoul Y, Thompson B, De Boer B (2018) Introducing Parselmouth: A Python interface to Praat. J Phon 71:1–15. 10.1016/j.wocn.2018.07.001

34. Boersma P, Weenink D Praat: doing phonetics by computer. Version 6.3.08; 2023. Available from: http://www.praat.org/
35. Baevski A, Zhou H, Mohamed A, Auli M (2020) wav2vec 2.0: A framework for self-supervised learning of speech representations. Adv Neural Inf Process Syst 33:12449–12460
36. Conneau A, Baevski A, Collobert R, Mohamed A, Auli M (2021) Unsupervised cross-lingual representation learning for speech recognition. In: Proc Interspeech. ; 2021. pp. 2426–2430. 10.21437/Interspeech.2021-329
37. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G et al (2019) PyTorch: An imperative style, high-performance deep learning library. Adv Neural Inf Process Syst 32:8024–8035
38. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A et al (2020) Transformers: State-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations; pp. 38–45
39. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B et al (2024) TRIPOD + AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ 385:e078378. 10.1136/bmj-2023-078378
40. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O et al (2011) Scikit-learn: Machine learning in Python. J Mach Learn Res 12:2825–2830
41. Cox DR (1958) The Regression Analysis of Binary Sequences. J R Stat Soc Ser B Methodol 20(2):215–232. 10.1111/j.2517-6161.1958.tb00292.x
42. Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20(3):273–297. 10.1007/BF00994018
43. Breiman L (2001) Random Forests. Mach Learn 45(1):5–32. 10.1023/A:1010933404324
44. Chen T, Guestrin C (2016) XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; pp. 785–794. 10.1145/2939672.2939785
45. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N et al (2010) Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology 21(1):128–138. 10.1097/EDE.0b013e3181c30fb2
46. Linardatos P, Papastefanopoulos V, Kotsiantis S, Explainable AI (2020) A Review of Machine Learning Interpretability Methods. Entropy 23(1):18. 10.3390/e23010018
47. Fagherazzi G, Fischer A, Ismael M, Despotovic V (2021) Voice for Health: The Use of Vocal Biomarkers from Research to Clinical Practice. Digit Biomark 5(1):78–88. 10.1159/000515346
48. Wang N, Yang D (2025) Speech emotion recognition using fine-tuned Wav2vec2.0 and neural controlleddifferential equations classifier. Shahamiri SR, editor. PLOS ONE. ;20(2):e0318297. 10.1371/journal.pone.0318297
49. Maji B, Nasreen S, Guha R, Routray A, Majumdar D, Poonam K (2024) Exploring Self-Supervised Models for Depressive Disorder Detection: A Study on Speech Corpora. In: 2024 46th Annual

International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). Orlando, FL, USA: IEEE

50. Kiyoshige E, Ogata S, Kwon N, Nakaoku Y, Hayashi C, Blaylock N et al (2025) Developing and testing AI-based voice biomarker models to detect cognitive impairment among community dwelling adults: a cross-sectional study in Japan. Lancet Reg Health - West Pac 59:101598. 10.1016/j.lanwpc.2025.101598
51. Herrmann BL, Mortsch F, Berg C, Weischer T, Mohr C, Mann K (2010) Acromegaly: A Cross-Sectional Analysis of the Oral and Maxillofacial Pathologies. Exp Clin Endocrinol Diabetes 119(01):9–14. 10.1055/s-0030-1255020
52. Karakis D, Aktas-Yilmaz B, Dogan A, Yetkin I, Bek B (2014) The bite force and craniofacial morphology in patients with acromegaly: A pilot study. Med Oral Patol Oral Cir Bucal e1–7. 10.4317/medoral.18984
53. Dostálová S, Šonka K, Šmahel Z, Weiss V, Marek J (2003) Cephalometric assessment of cranial abnormalities in patients with acromegaly. J Cranio-Maxillofac Surg 31(2):80–87. 10.1016/S1010-5182(02)00182-8
54. Robin J, Harrison JE, Kaufman LD, Rudzicz F, Simpson W, Yancheva M (2020) Evaluation of Speech-Based Digital Biomarkers: Review and Recommendations. Digit Biomark 4(3):99–108. 10.1159/000510820
55. Gold M, Amatniek J, Carrillo MC, Cedarbaum JM, Hendrix JA, Miller BB et al (2018) Digital technologies as biomarkers, clinical outcomes assessment, and recruitment tools in Alzheimer's disease clinical trials. Alzheimers Dement N Y N 4:234–242. 10.1016/j.trci.2018.04.003
56. Coravos A, Khozin S, Mandl KD (2019) Developing and adopting safe and effective digital biomarkers to improve patient outcomes. NPJ Digit Med 2(1):14. 10.1038/s41746-019-0090-4
57. Kourtis LC, Regele OB, Wright JM, Jones GB (2019) Digital biomarkers for Alzheimer's disease: the mobile/ wearable devices opportunity. NPJ Digit Med 2:9. 10.1038/s41746-019-0084-2
58. Macias Alonso AK, Hirt J, Woelfle T, Janiaud P, Hemkens LG (2024) Definitions of digital biomarkers: a systematic mapping of the biomedical literature. BMJ Health Care Inf 31(1):e100914. 10.1136/bmjhci-2023-100914
59. Sun YM, Wang ZY, Liang YY, Hao CW, Shi CH (2024) Digital biomarkers for precision diagnosis and monitoring in Parkinson's disease. NPJ Digit Med 7(1):218. 10.1038/s41746-024-01217-2
60. Babrak LM, Menetski J, Rebhan M, Nisato G, Zinggeler M, Brasier N et al (2019) Traditional Digit Biomarkers: Two Worlds Apart? Digit Biomark 3(2):92–102. 10.1159/000502000