# مجموعه های داده

## شرح مجموعه های داده

### NeuroVoz

مجموعه داده NeuroVoz یک مجموعه جامع از گفتار به زبان اسپانیایی کاستیلی است که برای تحقیقات بیماری پارکینسون طراحی شده است. این مجموعه شامل ضبط صدای ۵۴ فرد مبتلا به PD و ۵۸ فرد سالم (HC) است که همگی در یک محیط کنترل‌شده با استفاده از یک میکروفون هدست باکیفیت AKG C420 با فرکانس نمونه‌برداری ۴۴.۱ کیلوهرتز ضبط شده‌اند. یکی از نقاط قوت کلیدی این مجموعه داده، تنوع وظایف گفتاری است که توسط شرکت‌کنندگان انجام شده است، شامل:

* تولید ممتد پنج واکه اسپانیایی /a/، /e/، /i/، /o/، /u/
* یک تست DDK شامل تکرار سریع توالی /pa-ta-ka/
* ۱۶ جمله برای گوش دادن و تکرار
* یک مونولوگ خودانگیخته برای توصیف یک تصویر

این مجموعه داده با فراداده‌های گسترده‌ای غنی شده است، از جمله اطلاعات بالینی دقیق شامل مقیاس‌های UPDRS و H-Yو مهم‌تر از آن، ارزیابی‌های کیفیت صدای ذهنی که توسط یک متخصص با استفاده از مقیاس GRBAS انجام شده است. این ویژگی امکان مقایسه مستقیم بین تحلیل‌های محاسباتی و ارزیابی ادراکی بالینی را فراهم می‌کند.

### mPower

مطالعه mPower یک نقطه عطف در تحقیقات سلامت همراه (mobile health) است که داده‌ها را از راه دور از هزاران شرکت‌کننده از طریق یک اپلیکیشن اختصاصی آیفون جمع‌آوری کرده است. این مجموعه داده شامل ۵۸۲۶ شرکت‌کننده است که وظیفه صوتی را انجام داده‌اند، که از این تعداد ۱۰۸۷ نفر خود را مبتلا به PD معرفی کرده و ۵۵۸۱ نفر به عنوان گروه کنترل شرکت کرده‌اند. داده‌ها "در محیط واقعی" و بدون نظارت محقق جمع‌آوری شده‌اند که منعکس‌کننده شرایط دنیای واقعی است. فعالیت صوتی ساده و کوتاه بود: از شرکت‌کنندگان خواسته می‌شد واکه /a/ را تا ۱۰ ثانیه به صورت ممتد در میکروفون آیفون خود تولید کنند. مزیت کلیدی mPower مقیاس عظیم آن است که به لطف سهولت مشارکت از راه دور مبتنی بر اپلیکیشن امکان‌پذیر شده است. با این حال، این مزیت با محدودیت‌هایی همراه است، زیرا داده‌ها تنها از یک وظیفه صوتی ساده به دست آمده‌اند و فاقد کیفیت کنترل‌شده و بررسی‌های بالینی دقیق مجموعه‌داده‌های آزمایشگاهی هستند.

### مجموعه داده مطالعه ایتالیایی یا IPVS

این مطالعه که توسط Dimauro و همکارانش انجام شده، بر ارزیابی قابلیت فهم گفتار در بیماری پارکینسون متمرکز بوده است. این مجموعه داده شامل ضبط صدای ۲۸ بیمار PD، ۲۲ فرد سالمند سالم و ۱۵ فرد جوان سالم است که همگی به زبان ایتالیایی صحبت می‌کنند. پروتکل ضبط گسترده بود و از شرکت‌کنندگان می‌خواست چندین وظیفه را در یک اتاق ساکت با استفاده از میکروفون انجام دهند:

* خواندن یک متن متعادل از نظر آوایی
* خواندن لیست‌هایی از کلمات و عبارات متعادل از نظر آوایی
* تکرار هجاهای /pa/ و /ta/
* تولید ممتد هر پنج واکه ایتالیایی

ویژگی منحصربه‌فرد این پژوهش، روش‌شناسی آن است. محققان به جای تحلیل مستقیم ویژگی‌های آکوستیک، فایل‌های صوتی را به یک سیستم عمومی تبدیل گفتار به متن (STT) (Google STT) وارد کردند و نرخ خطای تشخیص کلمات را به عنوان معیاری برای قابلیت فهم گفتار اندازه‌گیری کردند. این مطالعه نشان داد که نرخ خطا به طور مداوم برای گروه PD در مقایسه با هر دو گروه کنترل سالم بالاتر بود، که این روش را به عنوان یک راهکار عینی برای اندازه‌گیری اختلال گفتار معرفی می‌کند.

### مجموعه داده مطالعه آرکانزاس یا UAMS

این مطالعه توسط Iyer و همکارانش، امکان‌سنجی استفاده از نمونه‌های صوتی تلفنی و با کیفیت پایین را برای تشخیص بیماری پارکینسون بررسی کرده است. این مجموعه داده از ۵۰ فرد مبتلا به PD (با تشخیص تایید شده توسط متخصص) و ۵۰ فرد سالم جمع‌آوری شده است که عمدتاً از مناطق روستایی و محروم از خدمات پزشکی در آرکانزاس بودند. شرکت‌کنندگان به سادگی با یک شماره تلفن تماس گرفته و یک پست صوتی دیجیتال از خود در حال تولید ممتد واکه /a/ به جا گذاشتند. این روش منجر به ضبط‌هایی با وضوح پایین و فرکانس نمونه‌برداری ۸ کیلوهرتز شد که کاهش کیفیت قابل توجهی نسبت به ضبط‌های آزمایشگاهی یا با گوشی هوشمند دارد.

## مقایسه مجموعه های داده

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| مطالعه | نام مجموعه داده | شرکت‌کنندگان (PD / HC / گروه‌های دیگر) | زبان | وظایف گفتاری | محیط ضبط | دستگاه / مشخصات ضبط | ویژگی‌های کلیدی تحلیل‌شده |
| مجموعه داده مطالعه Hlavnička و همکاران | Synthetic | ۲۲ PD، ۲۱ MSA، ۱۸ PSP، ۲۲ HC (مجموعاً ۸۳ نفر) | چکی | واکه‌های ممتد /A/ و /I/ (۳۳۷ ضبط) | اتاق با نویز محیطی کم | میکروفون هدست کاندنسر Opus 55 (Beyerdynamic)، وضوح ۱۶ بیت، فرکانس نمونه‌برداری ۴۸ کیلوهرتز | F0 مُدال، ارتعاشات زیرهارمونیک، SHR، stdF0، PSI، DSA، LIS، مقایسه با ۱۹ آشکارساز F0 |
| مجموعه داده مطالعه Dimauro و همکاران | Italian Parkinson’s Voice and Speech | ۲۸ PD، ۱۵ YHC، ۲۲ HEC (مجموعاً ۶۵ نفر) | ایتالیایی | ۲ بار خواندن متن متعادل از نظر آوایی، هجاهای /pa/ و /ta/، ۲ بار تولید هر واکه (/a/، /e/، /i/، /o/، /u/)، خواندن کلمات و عبارات متعادل از نظر آوایی | اتاق ساکت، بدون پژواک، گرم (حدود ۲۲ درجه سانتی‌گراد) | میکروفون نامشخص (فاصله ۱۵-۲۵ سانتی‌متر)، استفاده از سیستم Google Speech to Text برای تحلیل | قابلیت فهم گفتار (نرخ خطای STT)، مدت زمان گفتار، کلمات خوانده‌شده، کاراکتر بر ثانیه، پارامترهای Voxtester (طیف، فرمنت‌ها، نرخ DDK، افت شدت، SPL) |
| مجموعه داده مطالعه Iyer و همکاران | UAMS | ۴۰ PD، ۴۱ HC (مجموعاً ۸۱ نفر) | انگلیسی | واکه ممتد /a/ (حدود ۳ ثانیه، کوتاه‌شده به ۱.۵ ثانیه) | از راه دور، در محیط‌های طبیعی | پست صوتی دیجیتال از طریق خطوط تلفن، کدواژه ۱۶ بیتی، فرکانس نمونه‌برداری ۸ کیلوهرتز | ویژگی‌های سیگنال آکوستیک (F0، فرمنت‌ها، HNR، جیتر، شیمر)، بردارهای ویژگی طیفی (LPC، LAR، Cep، MFCC)، تصاویر طیف‌نگاره رنگی و سیاه‌وسفید (تحلیل‌شده توسط Inception V3 CNN) |
| مجموعه داده NeuroVoz (نسخه ۳.۰.۰) | NeuroVoz | ۵۴ PD، ۵۸ HC (مجموعاً ۱۱۲ نفر) | اسپانیایی کاستیلی | واکه‌های ممتد (/a/، /e/، /i/، /o/، /u/)، ۱۶ جمله گوش دادن و تکرار (LR)، تست DDK با /pa-ta-ka/، مونولوگ‌های خودانگیخته برای توصیف تصاویر | اتاق ساکت در روال بالینی (از نظر آکوستیک ایزوله نشده) | میکروفون هدست AKG® C420، پیش‌تقویت‌کننده، کارت صدای SoundBlaster® Live، فرکانس نمونه‌برداری ۴۴,۱۰۰ هرتز، SNR ۲۴.۳ دسی‌بل | مقیاس ادراکی GRBAS، فراداده‌های دموگرافیک/بالینی، رونویسی‌های دستی، ویژگی‌های کیفیت صدا (پرتوربیشن، نویز، پارامترهای لرزش) |
| مجموعه داده مطالعه mPower | mPower | ۱۰۸۷ PD (خوداظهاری)، ۵۵۸۱ HC (خوداظهاری) (مجموعاً ۹,۵۲۰ نفر با اشتراک‌گذاری گسترده موافقت کردند) | انگلیسی | تولید ممتد 'Aaaaah' (تا ۱۰ ثانیه)، فعالیت حافظه، فعالیت ضربه زدن، فعالیت راه رفتن، نظرسنجی‌ها (دموگرافیک، PDQ-8، MDS-UPDRS) | از راه دور، خودراهبر از طریق اپلیکیشن آیفون | میکروفون آیفون (فرکانس نمونه‌برداری مبهم: به عنوان "وضوح بالا" و همچنین "کیفیت پایین" در مقایسه با محیط‌های کنترل‌شده توصیف شده است) | فایل‌های صوتی از فعالیت صوتی، داده‌های شتاب‌سنج/ژیروسکوپ، داده‌های ضربه زدن، پاسخ‌های نظرسنجی |

## تأثیر روش‌های جمع‌آوری داده بر تشخیص بیماری پارکینسون

روش‌های مختلف جمع‌آوری داده به طور قابل توجهی بر مقیاس‌پذیری، کیفیت و کاربرد بالینی مدل‌های تشخیص PD تأثیر می‌گذارند، اما تحلیل‌های پیشرفته می‌توانند بسیاری از محدودیت‌های داده‌های با کیفیت پایین را برطرف کنند.

1. ضبط کنترل‌شده در مقابل ضبط از راه دور: به طور سنتی، محققان ضبط‌های کنترل‌شده و آزمایشگاهی (مانند مطالعه NeuroVoz) را برای اطمینان از داده‌های با کیفیت بالا و کم‌نویز ترجیح می‌دادند که تحلیل را ساده‌تر می‌کند. با این حال، این رویکرد پرهزینه است و دسترسی شرکت‌کنندگان را محدود می‌کند. روش‌های از راه دور، مانند استفاده از اپلیکیشن گوشی هوشمند (mPower) یا پست صوتی تلفنی (مطالعه آرکانزاس)، به طور چشمگیری مقیاس‌پذیری و دسترسی را افزایش می‌دهند. این امر امکان جمع‌آوری مجموعه داده‌های بسیار بزرگ‌تر و متنوع‌تر را فراهم می‌کند، از جمله از افرادی در جوامع روستایی یا محروم از خدمات پزشکی که در غیر این صورت ممکن است در تحقیقات شرکت نکنند.
2. کیفیت صدا و عملکرد مدل: یکی از نگرانی‌های اصلی در مورد جمع‌آوری داده از راه دور، کیفیت پایین‌تر صدا و نویز پس‌زمینه است. برخی مطالعات اولیه نتایج متناقضی در مورد قابلیت اطمینان ضبط‌های تلفنی برای استفاده بالینی نشان دادند. با این حال، مطالعه آرکانزاس به طور مؤثری نشان می‌دهد که روش تحلیلی از کیفیت بکر داده‌ها حیاتی‌تر است. محققان با تبدیل صدای تلفنی با وضوح پایین (۸ کیلوهرتز) به طیف‌نگاره و تحلیل آن‌ها با یک CNN قدرتمند، به عملکردی قابل مقایسه یا حتی برتر از مدل‌های آموزش‌دیده بر روی داده‌های با کیفیت بالا و ضبط‌شده در آزمایشگاه دست یافتند. این نشان می‌دهد که با رویکرد صحیح، ضبط‌های با کیفیت پایین که به راحتی قابل دستیابی هستند، برای تشخیص PD بسیار مؤثرند.
3. امکان‌سنجی برای کاربردهای بالینی: موفقیت مدل‌هایی که از ضبط‌های تلفنی و گوشی هوشمند استفاده می‌کنند، برای کاربردهای بالینی در دنیای واقعی بسیار حیاتی است. این روش‌ها ساده، مقرون‌به‌صرفه و غیرتهاجمی هستند و آن‌ها را برای توسعه ابزارهایی ایده‌آل می‌سازد که می‌توانند برای غربالگری اولیه توسط پزشکان عمومی یا برای نظارت از راه دور بر پیشرفت بیماری استفاده شوند. توانایی جمع‌آوری داده از خانه بیمار از طریق یک تماس تلفنی ساده یا یک اپلیکیشن، پیگیری علائم و پاسخ به درمان‌ها را بدون نیاز به مراجعات مکرر به کلینیک آسان‌تر می‌کند.

## آماده سازی داده ها

1. **آماده‌سازی اولیه داده‌ها (Data Preparation)** این مرحله شامل جمع‌آوری فایل‌های صوتی از منابع مختلف و سازماندهی آن‌ها در یک ساختار مشخص (برای مثال، پوشه‌های جداگانه برای بیماران و افراد سالم) است. در این مرحله، داده‌های مربوط به سن و جنسیت نیز از فایل‌های متادیتا (Metadata) استخراج شده و به هر فایل صوتی متصل می‌شوند. برای دیتاست‌های بزرگ و پیچیده، از کتابخانه‌هایی مانند **Pandas** برای مدیریت و پاک‌سازی داده‌های متادیتا استفاده می‌شود.
2. **افزایش داده (Data Augmentation)** یکی از چالش‌های اصلی در یادگیری ماشین، کمبود داده‌های کافی است. افزایش داده به ما کمک می‌کند تا با ایجاد نسخه‌های دستکاری‌شده از داده‌های موجود، تعداد نمونه‌ها را افزایش دهیم. تکنیک‌های رایج برای داده‌های صوتی شامل موارد زیر است:
   * **تغییر گام (Pitch Shift):** گام صدای اصلی را بدون تغییر سرعت آن بالا یا پایین می‌برد.
   * **تغییر سرعت (Time Stretching):** سرعت پخش فایل صوتی را بدون تغییر گام آن کم یا زیاد می‌کند.
   * **اضافه کردن نویز:** با اضافه کردن نویز سفید یا نویز محیطی به فایل صوتی، مدل در برابر صداهای مزاحم مقاوم‌تر می‌شود.
   * **تغییر بهره (Gain):** با تغییر حجم صدا، مدل به سطوح مختلف صدا حساسیت کمتری پیدا می‌کند.
3. **متعادل‌سازی دیتاست (Dataset Balancing)** در بسیاری از دیتاست‌های دنیای واقعی، تعداد نمونه‌های یک کلاس (مثلاً بیماران) کمتر از کلاس دیگر (افراد سالم) است. این عدم توازن می‌تواند باعث شود که مدل عملکرد ضعیفی در تشخیص کلاس اقلیت داشته باشد. برای حل این مشکل، از روش **Oversampling** استفاده می‌شود، که در آن نمونه‌های کلاس اقلیت به‌صورت تصادفی تکرار می‌شوند تا تعداد آن‌ها با کلاس اکثریت برابر شود.

## ویژگی ها

### روش‌های استخراج فیچر (Feature Extraction)

استخراج فیچر، فرآیند تبدیل داده‌های خام صوتی (موج صوتی) به نمایش‌های عددی است که اطلاعات مفیدی را در خود دارند و می‌توانند توسط مدل‌های یادگیری ماشین به راحتی پردازش شوند.

#### 1. فیچرهای طیفی (Spectral Features)

این فیچرها بر اساس تحلیل فرکانس صدا به دست می‌آیند و اطلاعاتی درباره محتوای فرکانسی سیگنال ارائه می‌دهند.

* **طیف‌نگار (Spectrogram):** یک نمایش بصری از طیف فرکانسی سیگنال است که با گذشت زمان تغییر می‌کند. محور افقی زمان، محور عمودی فرکانس و شدت رنگ، شدت (قدرت) فرکانس را نشان می‌دهد.
* **طیف‌نگار مل (Mel Spectrogram):** شبیه به طیف‌نگار است، اما مقیاس فرکانس آن بر اساس مقیاس ادراکی انسان (Mel) تنظیم شده است. این فیچر به‌طور گسترده در پردازش گفتار استفاده می‌شود.
* **ضریب‌های فرکانسی مل (Mel-frequency Cepstral Coefficients - MFCCs):** این فیچرها نمایانگر پاکت فرکانسی طیف صوتی هستند و ویژگی‌های مهمی از گفتار را به‌صورت فشرده ارائه می‌دهند. MFCCs به‌شدت در تشخیص گفتار و صدا کاربرد دارند.

#### 2. فیچرهای آکوستیک (Acoustic Features)

این فیچرها شامل معیارهای آکوستیکی سنتی هستند که ویژگی‌های فیزیکی صدا را توصیف می‌کنند و اغلب با استفاده از ابزارهایی مانند **Parselmouth** (که بر پایه Praat است) استخراج می‌شوند. این فیچرها به 23 دسته مختلف تقسیم می‌شوند:

* **کیفیت صدا (Voice Quality):** شامل میانگین فرکانس اصلی (F\_0), انحراف معیار F\_0, و نسبت هارمونیک به نویز (Harmonic-to-Noise Ratio - HNR) است که نشان‌دهنده میزان شفافیت صدا است.
* **لرزش صدا (Jitter):** به تغییرات دوره‌ای فرکانس اصلی از یک دوره به دوره بعدی اشاره دارد.
* **نوسان دامنه (Shimmer):** به تغییرات دوره‌ای دامنه سیگنال از یک دوره به دوره بعدی اشاره دارد.
* **فرمانت‌ها (Formants):** به قله‌های طیفی اشاره دارند که در اثر تشدید حفره‌های صوتی ایجاد می‌شوند. تحلیل فرمنت‌ها برای شناسایی ویژگی‌های واکه‌ها و تفاوت‌های آواشناسی بسیار مفید است.

این فرآیندها به مدل اجازه می‌دهند تا به جای تحلیل مستقیم موج صوتی خام، روی ویژگی‌های معنادار و فشرده‌سازی‌شده تمرکز کند، که این امر به بهبود عملکرد، کاهش زمان آموزش و افزایش دقت مدل کمک می‌کند.

# مدل

مدل شرح داده شده در مقاله، حدود 2میلیون پارامتر دارد که منجر به افزایش حجم مدل و همین طور، افزایش زمان آموزش آن می شود. برای همین موضوع، تغییراتی در آن انجام شد که منجر به کاهش تعداد پارامتر ها تا 600 هزار تا شد.

مدل پیاده‌سازی‌شده در این پروژه در طراحی بلوک نهایی طبقه بندی با مدل اصلی تفاوت دارد. در حالی که هر دو مدل از بخش‌های استخراج ویژگی مشابهی شامل CNN، Attention و LSTM استفاده می کنند، مدل فعلی از یک لایه گلوگاه (bottleneck) فشرده با ۱۲۸ نورون پس از الحاق بردار ویژگی‌ها استفاده می‌کند. این لایه به طور مؤثری ابعاد ویژگی‌ها را پیش از طبقه‌بندی نهایی کاهش می‌دهد. در مقابل، حجم بالای پارامتر در مدل مرجع به احتمال زیاد به دلیل استفاده از یک یا چند لایه تماماً متصل (Dense) با ابعاد بسیار بزرگ است که جزئیات آن در مشخصات مدل ذکر نشده و بخش عمده پارامترهای مدل را به خود اختصاص می‌دهد.

علاوه بر تفاوت در تعداد پارامترها، یک تفاوت ظریف در معماری جریان داده نیز میان دو مدل وجود دارد. مدل مرجع از یک رویکرد متوالی (sequential) پیروی می‌کند که در آن، خروجی مکانیزم توجه به عنوان ورودی برای لایه‌های LSTM عمل می‌کند (CNN -> Attention -> LSTM). اما در مدل پیاده‌سازی‌شده، این دو ماژول به صورت موازی (parallel) پردازش می‌شوند؛ به طوری که توالی خروجی از بلوک کانولوشنی به صورت همزمان هم به مکانیزم توجه و هم به لایه‌های LSTM ارسال می‌گردد (CNN -> (Attention, LSTM)). این رویکرد موازی به هر بخش اجازه می‌دهد تا به صورت مستقل، ویژگی‌های زمانی و بخش‌های مهم توالی را مستقیماً از خروجی CNN استخراج نماید.