Fırat Üniversitesi Deneysel ve Hesaplamalı Mühendislik Dergisi

FUJECE
Fırat University Journal of Experimental and Computational Engineering

Vision Based Transformer Tekniği İle İnsan Konuşmasından Nörolojik Bozuklukların Sınıflandırılması

Emel SOYLU[1*], Sema GÜL[2], Kübra ASLAN KOCA[3], Muammer TÜRKOĞLU[4], Murat TERZİ[5]

[1]Yazılım Mühendisliği Bölümü, Mühendislik Fakültesi, Samsun Üniversitesi, Samsun, Türkiye.
[2]Lisansüstü Enstitüsü, Nörobilim Bölümü, Ondokuz Mayıs Üniversitesi, Samsun, Türkiye.
[3]Yazılım Mühendisliği Bölümü, Mühendislik Fakültesi, Samsun Üniversitesi, Samsun, Türkiye.
[4]Nöroloji Bölümü, Tıp Fakültesi, Ondokuz Mayıs Üniversitesi, Samsun, Türkiye.
[1]emel.soylu@samsun.edu.tr, [2]sema.gul@omu.edu.tr, [3]kubraslantr@gmail.com,
[4]muammer.turkoglu@samsun.edu.tr, [5]mterzi@omu.edu.tr

Geliş Tarihi: 17.03.2024
Kabul Tarihi: 25.04.2024

Düzeltme Tarihi: 30.03.2024

doi: https://doi.org/10.62520/fujece.1454309
Araştırma Makalesi

Alıntı: E. Soylu, S. Gül, K. A. Koca, M. Türkoğlu ve M. Terzi, "Vision based transformer tekniği ile insan konuşmasından nörolojik bozuklukların sınıflandırılması", Fırat Üni. Deny. ve Hes. Müh. Derg., vol. 3, no 2, pp. 160-174, Haziran 2024.

Öz

Bu çalışmada, Parkinson hastalığı, Multipl Skleroz (MS), sağlıklı bireyler ve diğer kategoriler gibi farklı sağlık kategorilerinin insan konuşmasından yüksek doğrulukta sınıflandırılması için transformatör tabanlı sinir ağı yaklaşımını sunuyoruz. Bu yaklaşım, insan konuşmasının spektrogramlara dönüştürülmesinde yatmaktadır ve daha sonra bu spektrogramlar görsel görüntülere dönüştürülmektedir. Bu dönüşüm süreci, ağımızın çeşitli sağlık koşullarını belirten karmaşık ses desenlerini ve ince nüansları yakalamasını sağlar. Yaklaşımımızın deneysel doğrulaması, Parkinson hastalığı, MS, sağlıklı bireyler ve diğer kategoriler arasında yüksek performanslı sonuçlar vermiştir. Bu başarı, spektrogram analizi ve vision based transformer tekniği birleşimine dayanan yenilikçi, invaziv olmayan bir tanı aracı sunarak potansiyel klinik uygulamalar için kapıları açmaktadır.

Anahtar kelimeler: Nörolojik bozukluk sınıflandırması, Transformatör tabanlı sinir ağı, Ses sınıflandırması

[*] Yazışılan yazar

İntihal Kontrol: Evet – Turnitin
Şikayet: fujece@firat.edu.tr

Firat University Journal of Experimental
and Computational Engineering

FUJECE
Firat University Journal of Experimental and Computational Engineering

Vision Transformer Based Classification of Neurological Disorders from Human Speech

Emel SOYLU[1*], Sema GÜL[2], Kübra ASLAN KOCA[3], Muammer TÜRKOĞLU[4], Murat TERZİ[5]

[1,4] Department of Software Engineering, Faculty of Engineering, Samsun University, Samsun, Türkiye.
[2] Graduate Institute, Department of Neuroscience , Ondokuz Mayıs University, Samsun, Türkiye.
[3] Department of Software Engineering, Faculty of Engineering, Samsun University, Samsun, Türkiye.
[5]Department of Neurology, Faculty of Medicine, Ondokuz Mayıs University, Samsun, Türkiye.
[1]emel.soylu@samsun.edu.tr, [2]sema.gul@omu.edu.tr, [3]kubraslantr@gmail.com,
[4]muammer.turkoglu@samsun.edu.tr, [5]mterzi@omu.edu.tr

Received: 17.03.2024
Accepted: 25.04.2024

Revision: 30.03.2024

doi: https://doi.org/10.62520/fujece.1454309
Research Article

Citation: E. Soylu, S. Gül, K. A. Koca, M. Türkoğlu ve M. Terzi, “Vision transformer based classification of neurological disorders from human speech”, Firat Univ. Jour.of Exper. and Comp. Eng., vol. 3, no 2, pp. 160-174, June 2024.

Abstract

In this study, we introduce a transformative approach to achieve high-accuracy classification of distinct health categories, including Parkinson's disease, Multiple Sclerosis (MS), healthy individuals, and other categories, utilizing a transformer-based neural network. The cornerstone of this approach lies in the innovative conversion of human speech into spectrograms, which are subsequently transformed into visual images. This transformation process enables our network to capture intricate vocal patterns and subtle nuances that are indicative of various health conditions. The experimental validation of our approach underscores its remarkable performance, achieving exceptional accuracy in differentiating Parkinson's disease, MS, healthy subjects, and other categories. This breakthrough opens doors to potential clinical applications, offering an innovative, non-invasive diagnostic tool that rests on the fusion of spectrogram analysis and transformer-based models.

Keywords: Neurological disorder classification, Vision transformer, Audio classification

[*] Corresponding author

Plagiarism Checks: Yes – Turnitin
Complaints: fujece@firat.edu.tr

1. Introduction

Sound is one of the components that make up human perception in nature. The direction, intensity, and duration of sound play a major role in our understanding and interpretation of environmental events. Differences in sounds allow us to distinguish events. Sounds are generally composed of harmonic signals. The signal emitted in the air is received by the human ear, passed through certain neural processes in the brain, reaches the relevant auditory centers, and is interpreted [1]. Artificial intelligence techniques are used in sound analysis processes with the imitation of this mechanism.

Voice analysis of people is the subject of many fields of study. Speech signals are used as input sources in human-computer interaction to develop various applications such as automatic speech recognition, speech emotion recognition, gender and age recognition [2-4].

The human larynx functions in roles such as speaking, breathing, swallowing, coughing and has a complex functional structure. The coordination of these roles is very sensitive to being affected in individuals with neurological diseases. Sound problems arise because the larynx mechanism cannot meet the demand for sound due to functional or structural reasons. There are many factors that will affect the sound production mechanism. Speech habits, health problems, chronic diseases, habits, neurological disorders can be given as examples of factors that cause voice problems. In this study, neurological disease classification is made by voice analysis of patients diagnosed with neurological disease.

In the diagnosis of neurological diseases, the patient's history and physical examination usually come to the fore. Voice-related changes can often be overlooked [5]. Even if the patient does not have a complaint about the voice that can be expressed directly, the evaluation of the voice during the anamnesis or examination can make a significant contribution to the diagnosis of individuals with neurological diseases. With the effect of functions such as articulation and phonation during the speech, hypophonic, dysarthric, and ataxic sounds can contribute to the diagnosis. While hypophonic speech may suggest basal ganglia involvement with accompanying bradymia, it may suggest the involvement of the pyramidal pathway in a patient with first and second motor neuron findings. Dysarthric speech pattern suggests cerebellar involvement together with other threshold examination findings. A muffled speech pattern can be seen in motor neuron diseases, and speech problems up to motor or global aphasia can be seen in patients with cerebral cortical involvement. All these voice changes, together with the affected neurological system and other findings, provide important information about the diagnosis.

In the literature, there are studies based on computer-based processing of patients' voice data and early detection and diagnosis of diseases in health sciences. Abnormal condition detection by processing breath sound [4], [6], heart sound [7-12], knee joint sound processing for non-invasive diagnosis and monitoring of joint disorders such as osteoarthritis and chondromalacia [13], COVID-19 detection from cough, sound, breath sound, Alzheimer's detection from the speech process [14], Parkinson's detection [15] can be given as examples of sound processing studies in the health field [16-23].

The frequency spectrum of an audio signal can be expressed visually in the form of a spectrogram. The spectrogram can be constructed using an optical spectrometer, bank of band-pass filters, Fourier transform, or wavelet transform methods. Spectral representations are involved in classification or regression neural networks.

There are examples in the literature on converting audio signals to spectrograms and classifying them with artificial intelligence techniques. COVID-19 detection with lung breath sound [24], seizure detection from electroencephalography (EEG) signals [25], recognition of surrounding sounds, bird sound recognition [26], and emotion detection from community voice [27] are some examples of such studies.

The application of deep learning models for voice recognition in predicting vocal fold diseases related to neurological disorders has shown promise [28]. By leveraging spectrogram-based techniques, researchers have been able to develop AI tools for predicting vocal cord pathology in primary care settings, emphasizing the importance of spectrogram analysis in diagnosing voice-related issues [29]. The analysis of voice spectrograms plays a significant role in detecting, monitoring, and classifying neurological diseases based on voice characteristics. By utilizing advanced technologies like convolutional neural networks (CNN) and deep learning models, researchers are making strides in leveraging spectrogram data to improve the diagnosis and management of neurological conditions through voice analysis.

Transformer networks, initially designed for natural language processing (NLP), have found groundbreaking applications in various domains, including image classification. In image classification tasks, the primary objective is to categorize input images into predefined classes or labels [30]. Traditionally, CNNs have been the dominant choice for image classification due to their ability to capture spatial hierarchies within images. However, transformer networks have introduced a paradigm shift by leveraging attention mechanisms to process images in a non-sequential manner, making them highly effective in capturing global dependencies and relationships within image data [31-35].

Transformers find versatile applications in health data utilization. These applications span disease diagnosis by analyzing symptoms and medical histories, medical image processing such as segmentation and detection, drug discovery via genetic and molecular analysis, medical text processing for reports and records, biomedical natural language understanding, and health record management. Transformers offer a flexible framework for handling health data, showing great promise across various healthcare domains [36-39].

In the scope of this research, we employed transformer models to classify human voices, a task that sets the foundation for our investigation. Employing our proprietary dataset, we sourced audio recordings from both healthy individuals and those affected by conditions like Multiple Sclerosis and Parkinson's, introducing a distinctive dimension to our work. The collection process involved individuals with diagnosed conditions, underscoring our unique methodology. This unconventional approach significantly contributed to the ingenuity of our study, and because of this innovative dataset creation, we achieved remarkable levels of accuracy.

The upcoming sections of the paper encompass various aspects, including a review of related works, an in-depth exploration of the employed methodology, a detailed presentation of the dataset used, and an insightful discussion.

2. Relevant Work

In the literature, there are computer-based auxiliary studies for the detection of neurological diseases. In Table 1, the type of neurological disease, data types used in the diagnosis, method, and study years are given. This table provides information about various neurological disorders, the type of data used for diagnosis, the number of samples, diagnosis methods (such as Artificial Intelligence, Support Vector Machine, etc.), accuracy percentages, publication years, and references for each study. The disorders covered include Multiple Sclerosis (MS), Alzheimer's, and Parkinson's disease, along with the specific data types and methods used for diagnosis. As evident from the table, successful classification outcomes are achieved by training the language datasets obtained through speech features using artificial intelligence techniques.

The existing literature has bolstered our belief in the feasibility of classifying neurological diseases based on sound data. Contemporary literature explores the transformation of sound data into spectrograms and subsequent classification through deep learning techniques, although typically focusing on individual diseases. In contrast, our proposed study seeks to discern both MS and Parkinson's diseases from a dataset encompassing 12 patient categories (MS, Amyotrophic Lateral Sclerosis (ALS), Spinocerebellar Ataxia (SCA), Alzheimer's, Epilepsy, Parkinson's, Myasthenia Graves, Myelitis, Motor Aphasia, Psychological, Fiedreich Ataxia, Language Problem) and healthy individuals. A distinguishing feature of this study is the dataset's diverse range of disease types, setting it apart from prior research. We employ a larger patient cohort and solely utilize voice recordings for classification purposes. The utilization of everyday mobile phones for voice recording, as opposed to specialized devices, makes our approach practical, cost-effective, and distinct from prior dataset creation methodologies.

Table 1. Relevant work

Neurological Disorder	Data type to use in the diagnosis	Number of samples	Diagnosis method	Accuracy (%)	Year of Publication	Reference
MS	Conversation, demographics	65 patients, 66 healthy individuals	Artificial Intelligence	82	2022	[40]
MS	brain magnetic resonance images	168 lesion images obtained from 3 patients	Support Vector Machine	81.5	2010	[41]
Alzheimer	Speech Features	2033 audio recordings collected from 99 patients	Machine Learning	78.7	2019	[42]
Alzheimer	Speech Features	80 healthy, 13 early diagnoses, 5 patients	LDA classifier	85.7	2016	[43]
MS	surface electromyography (sEMG) signals	450-sentence speech collected from 3 sick individuals	CNN	81	2020	[44]
Parkinson	Pc-Gita (Vowel monologues, sentences,words, read text)	50 healthy, 50 Parkinson individuals	CNN	98.3	2020	[45]
Alzheimer	Speech Features	254 Alzheimer's and 250 Healthy individuals	Machine Learning	89.4	2020	[46]
Parkinson	Speech Features	91 subjects, 43 suffering from PD with each person on an average giving 5-6 different samples	CNN	89.15	2019	[47]
Parkinson, ALS	Speech Features	60 ALS, 60 Parkinson, 60 Healthy individuals	CNN	87	2020	[48]
Parkinson	Speech Features	120 speech samples from 20 Healthy, 28 Parkinson individuals	Generative Adversarial Network	90.5	2020	[49]
Parkinson	Speech Features	181 speakers, 1797 recordings from 3 different languages	Vision Transformer	78	2021	[50]
Parkinson	Gait	64468 gait data	Transformer	97.4	2022	[51]
Parkinson	Drawing	315 Healthy, 279 Parkinson sample	CNN	95.29	2023	[52]
MS, Parkinson	Speech Features	204 MS, 172 Parkinson, 212 Other, 94Healthy individuals	Vision Transformer	93.14	-	Our proposed research

3. Method

For many years the key point in audio analysis has been feature design and selection. In feature extraction, higher-order statistics of spectral center and spectral shape, zero crossing statistics, harmonics, fundamental frequency, and temporal explanations were used [53]. Today, feature extraction is done by deep networks. In this method, networks produce successful results when enough samples are used. The Fourier transform enables the representation of signals from the time domain into the frequency domain. This concept was introduced by Jean Baptiste Joseph Fourier, a French mathematician and physicist. Utilizing the Fourier Transform, the original time-based signal can be deconstructed into sinusoidal components, each possessing an amplitude, phase, and frequency. A waveform that appears complex in the time domain translates to a vertical line within the frequency domain. This concise depiction in the frequency domain serves to highlight essential frequencies. The Fourier Transform effectively dissects intricate time-based signals into distinct frequency constituents, simplifying comprehension. The transition from the frequency domain back to the time domain preserves all data, ensuring fidelity. Given that audio signals are dynamic and not static, their characteristics fluctuate over time. Consequently, attempting a single Fourier transform across an entire 5-minute lecture's speech would be impractical. Such an approach would yield indistinct data for analysis. Alternatively, the Fourier transform is applied to successive signal frames, introducing the concept of Short-Time Fourier Transform (STFT). This approach better accommodates the variable nature of audio signals and enhances the extraction of meaningful features for analysis. The human perception of sound intensity follows a logarithmic scale, emphasizing the significance of logarithmic amplitude. Calculating the logarithmic value can be achieved using the Librosa library. Librosa, a Python package tailored for music and audio analysis, encompasses essential components to construct information retrieval systems employed in audio analysis [54-56].

Utilizing spectrograms, which visually represent sound signals, and employing deep learning for classification purposes is crucial due to several reasons. Firstly, spectrograms enable the capture of complex patterns within audio data, facilitating comprehensive analysis. Secondly, by treating sound as images, deep learning models can extract relevant features using image processing techniques, enhancing classification accuracy. Additionally, spectrogram-based representations offer interpretable features, aiding in understanding the acoustic properties associated with different health conditions. Lastly, leveraging deep learning architectures designed for image classification tasks ensures compatibility and efficiency in health category classification. Overall, combining spectrograms and deep learning techniques presents a powerful approach for sound-based classification in healthcare, advancing diagnostic capabilities and patient care.

In this study, voice data from different neurological patients and healthy individuals were evaluated. Individuals are told a sample sentence and recorded. The collected data were labelled by the specialist physician. Diagnosis of diseases and affected neurological systems were recorded after neurological evaluation.

The dataset was created by eliminating the misleading ones in the obtained data. It is aimed to collect enough data to enable machine learning for each disease category. The steps of the method used in this study can be summarized as follows. The block diagram of the proposed system is given in Figure 1.

- Data collecting
- Data labelling
- Data elimination
- Extraction of spectrograms of audio signals
- Classification with vision transformer (ViT)

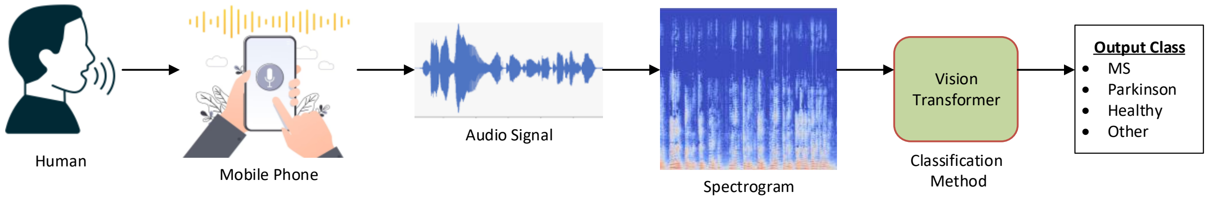


Figure 1. Patient voice classification block diagram

In the Colab environment, librosa library was used for the conversion of sound files to spectrograms, tensorflow library for convolutional neural network models, matplotlib library for graphical drawings, cv2 library for image cropping and resizing.

3.1. Dataset Description

The dataset used in this study was obtained by having patients diagnosed with MS, Parkinson and other neurological diseases (ALS, SCA, Alzheimer's, Epilepsy, Parkinson, Myasthenia Graves, Myelitis, Motor Aphasia, Psychological, Fiedreich Ataxia, Language Problem) and healthy individuals say a common sentence in Turkish. In the dataset, there were .wav audio files of 204 individuals with MS, 172 with Parkinson's, 212 with other neurological diseases, and 94 healthy individuals. The study group consisted of individuals between the ages of 18 and 65 who had been diagnosed with a neurological disease.

The dataset meticulously obtained and employed in this study deserves commendation for its unprecedented contribution to the field of healthcare and software development. It stands as a testament to the pioneering spirit of multidisciplinary research, bridging the realms of medicine and technology. This unique dataset, drawn from real patients and healthy individuals, represents a valuable resource that has paved the way for innovative and groundbreaking advancements. Its richness, authenticity, and comprehensiveness serve as the cornerstone of our transformative approach, allowing us to harness the power of artificial intelligence and machine learning for the early diagnosis and differentiation of neurological diseases. In Figure 2, the process of obtaining the spectrogram graph from the audio file and cutting the image is shown visually, respectively. By applying this process to each sound file, spectrogram images are obtained. The details of the dataset are given in Table 2. 85% of these images are used to create a classification model and 15% is used to test the accuracy of the model. The training and test data in the dataset are randomly determined.

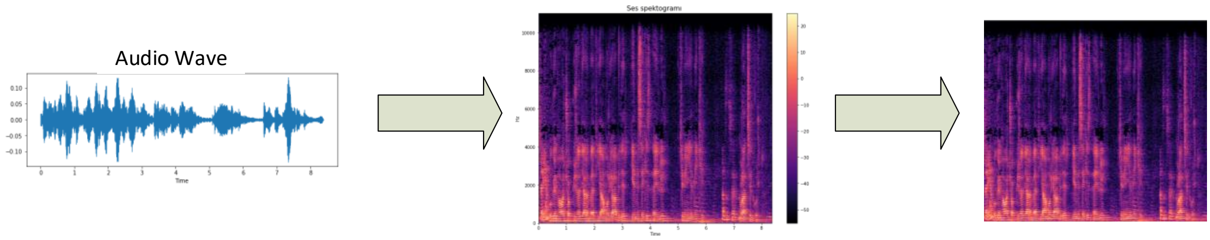


Figure 2. Process of obtaining spectrogram from audio file

Table 2. Details of the dataset

Class	Number of samples	Train data	Test Data
Healthy	94	80	14
MS	204	174	30
Parkinson	172	146	26
Other	212	180	32
Total	682	580	102

Google Colaboratory, shortly Colab, was used for image classification with dataset editing and transfer-based deep learning technique. Colab is a product offered by Google Research. It is particularly suitable for machine learning, data analysis, and education. With Colab, people can write and execute Python code through the browser.

3.2.Transformers

Unlike CNNs that process images pixel by pixel, transformer networks divide images into N patches fixed-size patches. This process is called linear embedding of patches (Elin). Each image patch xi is linearly embedded into a lower-dimensional space using the Elin operation enabling the network to capture essential features given in Eq.1.

$$z_i = Elin(x_i) \tag{1}$$

Transformer networks, originally developed for sequence data, don't inherently possess spatial information. To address this, positional encodings are introduced, allowing the model to understand the relative positions of patches. To incorporate positional information, positional encodings are added to the token embeddings as given in Eq.2.

$$z_{i\_pos} = z_i + Epos(i) \tag{2}$$

The self-attention mechanism is a key component of transformers. It allows each patch to attend to all other patches, capturing long-range relationships and enabling the model to recognize complex patterns. The self-attention mechanism calculates attention scores and output embeddings for each pair of tokens (i, j) as given in Eq.3.

$$Attention\left(z_{i\_pos}, z_{j\_pos}\right) = Softmax\frac{(z_{i\_pos} \cdot z_{j\_pos})^T}{\sqrt{d_k}} \cdot z_{j\_pos} \tag{3}$$

The transformer encoder processes the embedded patches along with positional encodings. Multiple layers of encoders capture hierarchical features and generate context-aware representations. Multi-Head Self-Attention combines multiple attention heads to capture different relationships. The equation of this process is given in Eq.4. Each Head_k operates similarly to the self-attention mechanism but with different learned weight matrices. Each Head_k operates similarly to the self-attention mechanism but with different learned weight matrices.

$$MultiHead\left(z_{i\_pos}\right) = Concat\left(Head_{1\left(z_{i\_pos}\right)}, \ldots., Head_{h\left(z_{i\_pos}\right)}\right) \cdot W\_o \tag{4}$$

The Transformer Encoder processes the output of the multi-head self-attention and combines it with the original input. The equation of this process is given in Eq.5. LayerNorm performs layer normalization, and the output is added to the original token embedding.

$$Output_i = LayerNorm(MultiHead\left(z_{i\_pos}\right) + z_{i\_pos}) \tag{5}$$

Once the patch representations are processed, a global classification token is added. This token aggregates information from all patches and contributes to the final classification decision. The aggregation function combines all the token embeddings into a single global token as given in Eq. 6.

$$Global\_Token = Aggregation(z_{1\_pos}, \ldots, z_{N\_pos}) \tag{6}$$

The global token's representation is fed into the classification layer, which maps it to the respective classes as given in Eq. 7. Here, W_cls represents the weight matrix for the classification layer

$$Class\_Logits = Global\_Token \cdot W_{cls} \tag{7}$$

The softmax activation function is applied to the class logits to obtain class probabilities as given in Eq. 8. We chose to use the softmax activation function to obtain class probabilities because softmax ensures that the output probabilities sum up to 1, which is desirable for interpreting the output as probabilities of different classes. Additionally, softmax normalizes the logits, making them more interpretable and suitable for multi-class classification tasks. Softmax is more suitable for obtaining class probabilities at the output layer.

$$Class\_Probabilities = Softmax(Class\_Logits) \tag{8}$$

The model is trained using labelled image data and a loss function. Commonly used loss functions include categorical cross-entropy. After training, the model's accuracy is evaluated on unseen test data. Transformer networks' success in image classification highlights their adaptability to diverse data types and tasks beyond

NLP. They have demonstrated state-of-the-art performance on various benchmark datasets and have contributed to pushing the boundaries of image understanding. The architecture of model is given in Figure 3. The illustration is inspired from [57].

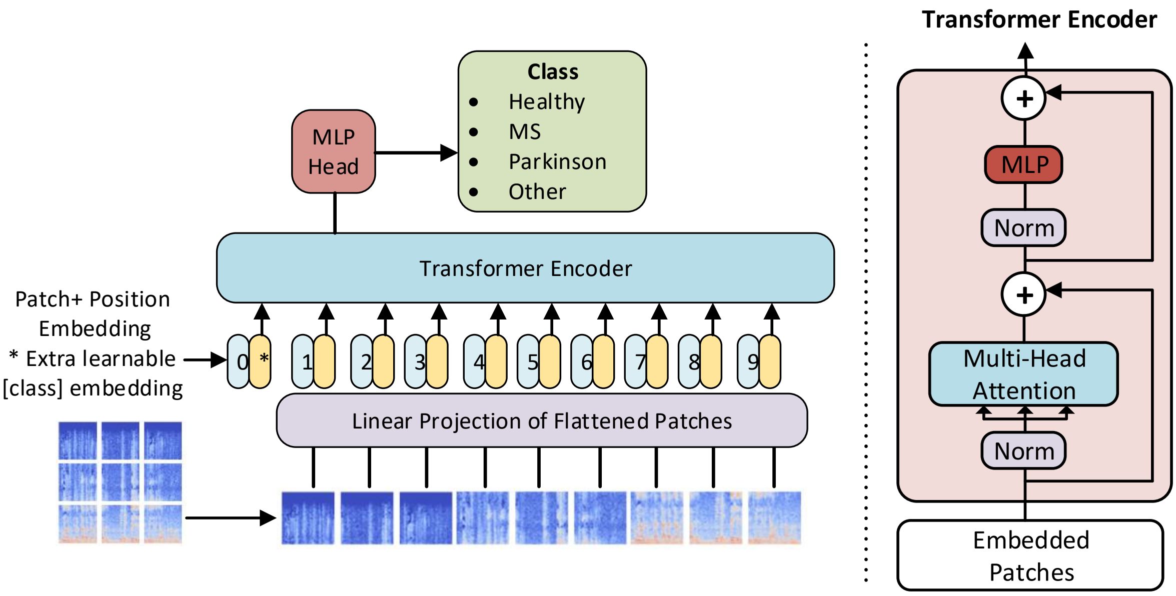


Figure 3. Model overview

In our research, we conducted a retraining process on Google's 'vit-base-patch16-224-in21k' model. The ViT-Base model comprises 12 layers with a hidden size of 768, an MLP size of 3072, and incorporates 12 attention heads, totaling 86 million parameters [57]. Utilizing the Adam optimizer and setting a learning rate of 1e-6, along with adjusting hyperparameters after 100 epochs, we managed to decrease the training loss to 0.05. The epoch count was decided after numerous iterations, reaching the desired error level as the determining factor. The training progress outcomes are depicted in Figure 4.

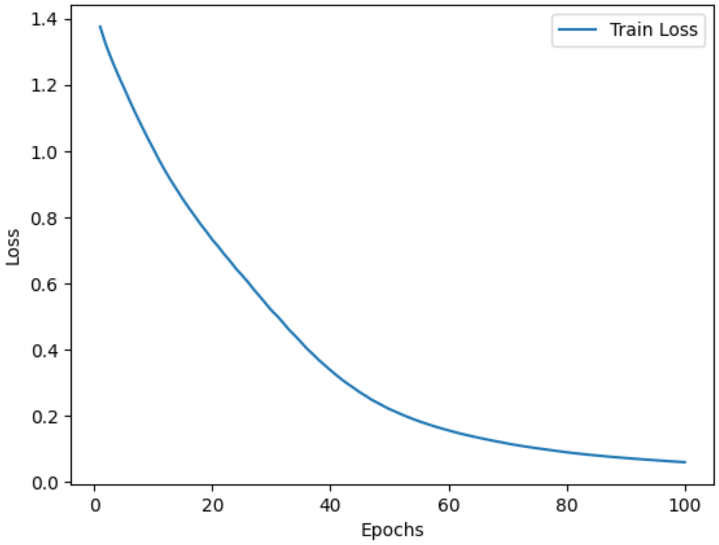
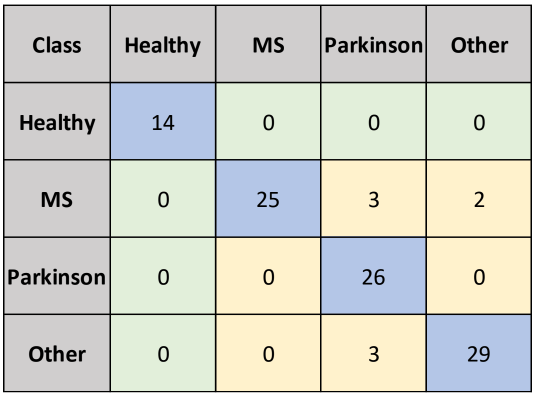


Figure 4. Train loss curve

4. Results

In this study, a dataset was created from the recorded data by having 682 individuals with neurological diseases say an example sentence. By transforming the audio data in the created dataset into spectrogram images, the classification of neurological diseases is provided by vision transformer-based learning. The confusion matrix of experiments is given in Figure 5. As a result of experiments with test data, an 92.15% success rate was obtained.



Class	Healthy	MS	Parkinson	Other
Healthy	14	0	0	0
MS	0	25	3	2
Parkinson	0	0	26	0
Other	0	0	3	29

Figure 5. Confusion Matrix

Table 3 contains various metrics used to evaluate the performance of a classification model. n truth (True Count) is the number of actual data points for each class. n classified (Classified Count) is the number of data points correctly classified by the model for each class. Accuracy measures the ratio of correct predictions made by the model to the total number of data points. This metric assesses the overall performance of the model. Precision indicates the proportion of positive predictions that are true positives. It measures how accurate the model's positive predictions are for a class. Recall measures how many of the true positive examples were correctly predicted. It assesses the model's ability to capture true positives. For instance, the recall for the "Parkinson" class is 94%, indicating that most of the true positives for this class were correctly predicted. The F1 score is a metric that balances precision and recall. Ideally, you want to achieve a high F1 score with both high precision and recall. As can be seen from the table, "Healthy" and "Parkinson" classes have high accuracy and F1 scores. Out of 102 test inputs, 94 were correctly predicted, resulting in an overall accuracy of 92.15%.

Table 3. Results for ViT

		Truth data					
		Healthy	**MS**	**Parkinson**	**Other**	**Classification overall**	**F1 Score**
Classifier results	**Healthy**	14	0	0	0	14	100%
	MS	0	25	3	2	30	83%
	Parkinson	0	0	26	0	26	100%
	Other	0	0	3	29	32	91%
	Truth overall	14	25	32	31	102	
	Recall	100%	100%	81.25%	93.55%		
	Overall accuracy	92.15%					

We also conducted testing on the dataset using various other architectures, resulting in the outcomes presented in Table 3. DenseNet121 achieved an accuracy of 67.91%. DenseNet201 achieved an accuracy of 64.93%. Xception achieved an accuracy of 73.88%. InceptionV3 achieved an accuracy of 72%. MobileNet achieved an accuracy of 73.13%. EfficientNetB0 achieved an accuracy of 80.6%. EfficientNetB0 architecture showed strong performance with an accuracy of 80.6%. EfficientNetV2B3 achieved an accuracy of 76.87%. Vision Transformer achieved an accuracy of 92.15%. The Vision Transformer architecture performed exceptionally well with an accuracy of 92.15%. This indicates that it correctly classified a vast majority of the test data, demonstrating its effectiveness for the task.

Table 4. Accuracy results for other deep learning techniques

Architecture	Accuracy
DenseNet121	67.91 %
DenseNet201	64.93 %
Xception	73.88 %
InceptionV3	72 %
MobileNet	73.13 %
EficientNetB0	80.6 %
EfficientNetV2B3	76.87 %
Vision Transformer	92.15%

5. Discussion

In recent years, the intersection of artificial intelligence and healthcare has opened up exciting possibilities for the early detection and diagnosis of various medical conditions. One particularly promising area of research is the classification of neurological diseases through the analysis of audio data. This innovative approach holds immense potential in revolutionizing the field of healthcare, especially in the realm of neurological disorders.

Neurological diseases encompass a wide range of disorders, including but not limited to Parkinson's disease, Multiple Sclerosis (MS), Alzheimer's disease, and more. Early detection and accurate diagnosis of these conditions are pivotal for improving patient outcomes and enhancing the quality of life for individuals affected by these diseases. Audio-based neurological disease classification plays a crucial role in achieving early detection. By analyzing the unique vocal patterns and speech characteristics of patients, machine learning models can identify subtle deviations that might indicate the presence of neurological disorders. This non-invasive approach can significantly reduce the time between symptom onset and diagnosis, enabling timely medical intervention and treatment.

One of the most compelling aspects of audio-based classification is its potential to democratize healthcare. Access to specialized medical facilities and experts can be limited in many regions, particularly in rural or underserved areas. Audio-based diagnostic tools can be easily distributed and utilized remotely, bridging the gap between patients and healthcare resources. Patients can record their speech and vocal samples in the comfort of their homes, making it easier to monitor their health and share data with healthcare providers. This accessibility not only reduces the burden on healthcare systems but also empowers individuals to take a more active role in managing their health. Audio-based neurological disease classification offers an objective and quantitative approach to diagnosis. Traditional diagnostic methods may rely on subjective assessments or a series of clinical tests, which can be prone to human error or bias. In contrast, machine learning models process audio data with consistency, providing reliable and reproducible results.

These models can analyze a multitude of features within speech data, detecting subtle changes that may escape human observation. As a result, healthcare professionals can make more informed decisions, leading to improved patient care. The field of audio-based neurological disease classification is still evolving, presenting exciting opportunities for further research and innovation. The development of advanced machine learning algorithms and the integration of state-of-the-art technologies, such as Vision Transformers, promise even greater accuracy and specificity in disease classification.

Additionally, collaborative efforts between researchers, clinicians, and technology experts are crucial for advancing this field. As we continue to refine and expand our understanding of neurological disease markers in audio data, we pave the way for novel diagnostic tools that can benefit millions of individuals worldwide.

6. Conclusions

In this study, we propose a transformative approach that aims to achieve highly accurate classification within different health categories, including Parkinson's disease, Multiple Sclerosis (MS), healthy individuals, and other categories. We utilize a transformer-based neural network as the basis of our approach. The most important innovation lies in the conversion of human speech into spectrograms, which are then converted into visual images. This transformation process allows our neural network to capture complex sound patterns and subtle nuances that indicate various health conditions. Experimental validation of our approach yields impressive results. It demonstrates exceptional accuracy in distinguishing between Parkinson's disease, MS, healthy subjects, and other categories. This breakthrough offers an innovative and non-invasive diagnostic tool that combines spectrogram analysis with transformer-based models, offering promising prospects for potential clinical applications.

We also conducted comparative experiments with various other deep learning architectures. These experiments demonstrate the superiority of ViT in achieving an accuracy of 92.15%, outperforming other architectures in correctly classifying health categories. We are currently continuing data collection and actively working to broaden the range of classes while striving to attain higher levels of accuracy in our forthcoming research endeavors.

Our research marks a promising step towards leveraging advanced machine learning techniques, specifically Vision Transformers, for effective health category classification based on sound patterns. These results have significant potential to improve diagnostic capabilities and contribute to non-invasive medical evaluations. Our research presents a non-invasive diagnostic tool capable of significantly enhancing the accuracy and efficiency of clinical diagnoses, particularly for conditions like Parkinson's disease and Multiple Sclerosis. By effectively classifying health categories based on sound patterns, healthcare professionals are empowered to make more informed decisions regarding patient care and treatment plans. Moreover, our method opens new horizons in remote monitoring and telemedicine applications. The ability to convert human speech into spectrograms and analyze them using transformer-based neural networks enables the development of smartphone apps or similar devices for recording and remotely analyzing speech patterns. This innovation can revolutionize healthcare delivery by facilitating timely assessments of patients' health statuses without the need for in-person visits. Additionally, our approach holds promise for early detection and prevention efforts. By capturing subtle nuances in sound patterns associated with various health conditions, it has the potential to identify early signs of Parkinson's disease, Multiple Sclerosis, and other ailments. Such early detection can prompt timely interventions, ultimately leading to improved patient outcomes and reduced healthcare burdens. Our research serves as a catalyst for further advancements in machine learning techniques within healthcare. It lays the groundwork for future studies to expand the range of classes beyond Parkinson's disease and Multiple Sclerosis, and to enhance the accuracy of classification algorithms. Through continued research and development, our approach stands to revolutionize healthcare practices, offering innovative solutions to pressing medical challenges.

7. Acknowledgement

This study attracted great attention at the 58th National Neurology Congress in Turkey and was awarded the second prize in the oral presentation category.

8. Author Contribution Statement

In the study, Murat Terzi, Emel Soylu, Sema Gül contributed to the creation of the idea, design, and literature review; Emel Soylu, Kübra Arslan Koca, Muammer Türkoğlu contributed to the analysis of the results, the provision of materials, and the review of the results.

9. Ethics Committee Approval and Conflict of Interest

The study protocol for this dataset was approved by the Ondokuz Mayıs University Clinical Research Ethics Committee (2022-545/2023). Written informed consent form was obtained from the address in the working environment and patient contents were extracted to ensure anonymity.

10. References

[1] B. Karasulu, “Çoklu ortam sistemleri için siber güvenlik kapsamında derin öğrenme kullanarak ses sahne ve olaylarının tespiti,” Acta INFOLOGICA, vol. 3, no. 2, pp. 60–82, 2019.

[2] A. Tursunov, J. Y. Choeh, and S. Kwon, “Age and gender recognition using a convolutional neural network with a specially designed multi-attention module through speech spectrograms,” Sensors, vol. 21, no. 17, p. 5892, 2021.

[3] M. Vacher, J.-F. Serignat, and S. Chaillol, “Sound classification in a smart room environment: an approach using GMM and HMM methods,” in The 4th IEEE Conference on Speech Technology and Human-Computer Dialogue (SpeD 2007), Publishing House of the Romanian Academy (Bucharest), 2007, vol. 1, pp. 135–146.

[4] J. Acharya and A. Basu, “Deep neural network for respiratory sound classification in wearable devices enabled by patient specific model tuning,” IEEE Trans. Biomed. Circuits Syst., vol. 14, no. 3, pp. 535–544, 2020.

[5] G. Woodson, “Management of neurologic disorders of the larynx,” Ann. Otol. Rhinol. \& Laryngol., vol. 117, no. 5, pp. 317–326, 2008.

[6] A. Abushakra and M. Faezipour, “Acoustic signal classification of breathing movements to virtually aid breath regulation,” IEEE J. Biomed. Heal. informatics, vol. 17, no. 2, pp. 493–500, 2013.

[7] E. Soares, P. Angelov, and X. Gu, “Autonomous learning multiple-model zero-order classifier for heart sound classification,” Appl. Soft Comput., vol. 94, p. 106449, 2020.

[8] Z. Dokur and T. Ölmez, “Heart sound classification using wavelet transform and incremental self-organizing map,” Digit. Signal Process., vol. 18, no. 6, pp. 951–959, 2008.

[9] M. Tschannen, T. Kramer, G. Marti, M. Heinzmann, and T. Wiatowski, “Heart sound classification using deep structured features,” in 2016 Computing in Cardiology Conference (CinC), 2016, pp. 565–568.

[10] P. Langley and A. Murray, “Heart sound classification from unsegmented phonocardiograms,” Physiol. Meas., vol. 38, no. 8, p. 1658, 2017.

[11] Z. Ren, N. Cummins, V. Pandit, J. Han, K. Qian, and B. Schuller, “Learning image-based representations for heart sound classification,” in Proceedings of the 2018 international conference on digital health, 2018, pp. 143–147.

[12] M. Deng, T. Meng, J. Cao, S. Wang, J. Zhang, and H. Fan, “Heart sound classification based on improved MFCC features and convolutional recurrent neural networks,” Neural Networks, vol. 130, pp. 22–32, 2020.

[13] K. S. Kim, J. H. Seo, J. U. Kang, and C. G. Song, “An enhanced algorithm for knee joint sound classification using feature extraction based on time-frequency analysis,” Comput. Methods Programs Biomed., vol. 94, no. 2, pp. 198–206, 2009.

[14] I. Vigo, L. Coelho, and S. Reis, “Speech-and language-based classification of alzheimer’s disease: a systematic review,” Bioengineering, vol. 9, no. 1, p. 27, 2022.

[15] J. Rusz et al., “Speech biomarkers in rapid eye movement sleep behavior disorder and parkinson disease,” Ann. Neurol., vol. 90, no. 1, pp. 62–75, 2021.

[16] K. K. Lella and A. Pja, “Automatic diagnosis of covıd-19 disease using deep convolutional neural network with multi-feature channel from respiratory sound data: cough, voice, and breath,” Alexandria Eng. J., vol. 61, no. 2, pp. 1319–1334, 2022.

[17] M. Faezipour and A. Abuzneid, “Smartphone-based self-testing of covıd-19 using breathing sounds,” Telemed. e-Health, vol. 26, no. 10, pp. 1202–1205, 2020.

[18] N. Melek Manshouri, “Identifying covıd-19 by using spectral analysis of cough recordings: a distinctive classification study,” Cogn. Neurodyn., vol. 16, no. 1, pp. 239–253, 2022.

[19] N. Sharma et al., "Coswara--a database of breathing, cough, and voice sounds for covıd-19 diagnosis," arXiv Prepr. arXiv2005.10548, 2020.

[20] A. Tena, F. Clarià, and F. Solsona, "Automated detection of covıd-19 cough," Biomed. Signal Process. Control, vol. 71, p. 103175, 2022.

[21] L. Kranthi Kumar and P. J. A. Alphonse, "COVID-19 disease diagnosis with light-weight CNN using modified MFCC and enhanced GFCC from human respiratory sounds," Eur. Phys. J. Spec. Top., pp. 1–18, 2022.

[22] M. Kuluozturk et al., "DKPNet41: directed knight pattern network-based cough sound classification model for automatic disease diagnosis," Med. Eng. \& Phys., p. 103870, 2022.

[23] T. Nguyen and F. Pernkopf, "Lung sound classification using co-tuning and stochastic normalization," IEEE Trans. Biomed. Eng., 2022.

[24] T. Tuncer, E. Akbal, E. Aydemir, S. B. Belhaouari, and S. Dogan, "A novel local feature generation technique based sound classification method for covid-19 detection using lung breathing sound," Eur. J. Tech., vol. 11, no. 2, pp. 165–174, 2021.

[25] G. C. Jana, R. Sharma, and A. Agrawal, "A 1D-CNN-spectrogram based approach for seizure detection from EEG signal," Procedia Comput. Sci., vol. 167, pp. 403–412, 2020.

[26] J. Xie, K. Hu, M. Zhu, J. Yu, and Q. Zhu, "Investigation of different CNN-based models for improved bird sound classification," IEEE Access, vol. 7, pp. 175353–175361, 2019.

[27] V. Franzoni, G. Biondi, and A. Milani, "Crowd emotional sounds: spectrogram-based analysis using convolutional neural network.," in SAT@ SMC, pp. 32–36, 2019.

[28] H. Hu et al., "Deep learning application for vocal fold disease prediction through voice recognition: preliminary development study," J. Med. Internet Res., 2021.

[29] E. C. Compton et al., "Developing an artificial ıntelligence tool to predict vocal cord pathology in primary care settings," Laryngoscope, 2022.

[30] A. Vaswani et al., "Attention is all you need," Adv. Neural Inf. Process. Syst., vol. 30, 2017.

[31] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," ACM Comput. Surv., vol. 54, no. 10s, pp. 1–41, 2022.

[32] S. Bhojanapalli, A. Chakrabarti, D. Glasner, D. Li, T. Unterthiner, and A. Veit, "Understanding robustness of transformers for image classification," in Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 10231–10241.

[33] K. S. Kalyan, A. Rajasekharan, and S. Sangeetha, "Ammus: A survey of transformer-based pretrained models in natural language processing," arXiv Prepr. arXiv2108.05542, 2021.

[34] J. Beal, E. Kim, E. Tzeng, D. H. Park, A. Zhai, and D. Kislyuk, "Toward transformer-based object detection," arXiv Prepr. arXiv2012.09958, 2020.

[35] Z. Shao et al., "Transmil: Transformer based correlated multiple instance learning for whole slide image classification," Adv. Neural Inf. Process. Syst., vol. 34, pp. 2136–2147, 2021.

[36] F. Shamshad et al., "Transformers in medical imaging: a survey," Med. Image Anal., p. 102802, 2023.

[37] A. Hatamizadeh et al., "Unetr: Transformers for 3d medical image segmentation," in Proceedings of the IEEE/CVF winter conference on applications of computer vision, 2022, pp. 574–584.

[38] Z. Liu, Q. Lv, Z. Yang, Y. Li, C. H. Lee, and L. Shen, "Recent progress in transformer-based medical image analysis," Comput. Biol. Med., p. 107268, 2023.

[39] Z. Liu and L. Shen, "Medical image analysis based on transformer: A review," arXiv Prepr. arXiv2208.06643, 2022.

[40] E. Svoboda, T. Boril, J. Rusz, T. Tykalova, D. Horakova, C. Guttman, K. B. Blagoev, H. Hatabu and V. Valtchinov, "Assessing clinical utility of Machine Learning and Artificial Intelligence approaches to analyze speech recordings in Multiple Sclerosis: A Pilot Study," arXiv Prepr. arXiv2109.09844, 2021.

[41] D. Yamamoto et al., "Computer-aided detection of multiple sclerosis lesions in brain magnetic resonance images: False positive reduction scheme consisted of rule-based, level set method, and support vector machine," Comput. Med. Imaging Graph., vol. 34, no. 5, pp. 404–413, 2010.

[42] F. Haider, S. De La Fuente, and S. Luz, "An assessment of paralinguistic acoustic features for detection of Alzheimer's dementia in spontaneous speech," IEEE J. Sel. Top. Signal Process., vol. 14, no. 2, pp. 272–281, 2019.

[43] J. Weiner, C. Herff, and T. Schultz, "Speech-based detection of alzheimer's disease in

conversational german.," in Interspeech, 2016, pp. 1938–1942.

[44] A. Kapur, U. Sarawgi, E. Wadkins, M. Wu, N. Hollenstein, and P. Maes, "Non-ınvasive silent speech recognition in multiple sclerosis with dysphonia," Proc. Mach. Learn. Heal. NeurIPS Work., pp. 25–38, 2020.

[45] L. Zahid et al., "A spectrogram-based deep feature assisted computer-aided diagnostic system for Parkinson's disease," IEEE Access, vol. 8, pp. 35482–35495, 2020.

[46] L. Liu, S. Zhao, H. Chen, and A. Wang, "A new machine learning method for identifying Alzheimer's disease," Simul. Model. Pract. Theory, vol. 99, p. 102023, 2020.

[47] A. Johri, A. Tripathi, and others, "Parkinson disease detection using deep neural networks," in 2019 Twelfth international conference on contemporary computing (IC3), 2019, pp. 1–4.

[48] B. N. Suhas et al., "Speech task based automatic classification of ALS and Parkinson's Disease and their severity using log Mel spectrograms," in 2020 international conference on signal processing and communications (SPCOM), 2020, pp. 1–5.

[49] Z.-J. Xu, R.-F. Wang, J. Wang, and D.-H. Yu, "Parkinson's disease detection based on spectrogram-deep convolutional generative adversarial network sample augmentation," IEEE Access, vol. 8, pp. 206888–206900, 2020.

[50] D. Hemmerling et al., "Vision transformer for parkinson's disease classification using multilingual sustained vowel recordings."

[51] H.-J. Sun and Z.-G. Zhang, "Transformer-based severity detection of parkinson's symptoms from gait," in 2022 15th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), 2022, pp. 1–5.

[52] S. M. Abdullah et al., "Deep transfer learning based parkinson's disease detection using optimized feature selection," IEEE Access, vol. 11, pp. 3511–3524, 2023.

[53] L. Wyse, "Audio spectrogram representations for processing with convolutional neural networks," vol. 1, no. 1, pp. 37–41, 2017.

[54] F. Ye and J. Yang, "A deep neural network model for speaker identification," Appl. Sci., vol. 11, no. 8, p. 3603, 2021.

[55] "Stft." [Online]. Available: https://musicinformationretrieval.com/stft.html.

[56] B. Li, "On identity authentication technology of distance education system based on voiceprint recognition," in Proceedings of the 30th Chinese Control Conference, 2011, pp. 5718–5721.

[57] A. Dosovitskiy et al., "An image is worth 16x16 words: transformers for image recognition at scale. arxiv 2020," arXiv Prepr. arXiv2010.11929, 2010.