Engineering Applications of Artificial Intelligence 149 (2025) 110558

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

Speech signal-based accurate neurological disorders detection using convolutional neural network and recurrent neural network based deep network

Emel Soylu [a,*], Sema Gül [b], Kübra Aslan Koca [a], Muammer Türkoğlu [a], Murat Terzi [c], Abdulkadir Şengür [d]

[a] Samsun University, Faculty of Engineering and Natural Sciences, Department of Software Engineering, Samsun, Turkey
[b] Ondokuz Mayıs University, Graduate Institute, Department of Neuroscience, Samsun, Turkey
[c] Ondokuz Mayıs University, Faculty of Medicine, Department of Neurology, Samsun, Turkey
[d] Firat University, Department of Electrical and Electronic Engineering, Faculty of Technology, Elazığ, Turkey

ARTICLE INFO

Keywords:
Deep learning
Audio classification
Neurological diseases
Recurrent neural network
Gated recurrent unit

ABSTRACT

Neurological diseases often manifest in subtle alterations to the human voice due to damage in the sound-related regions of the brain. Leveraging advancements in artificial intelligence (AI) technologies, computers can discern minute variations in sound imperceptible to the human ear, enabling rapid and precise diagnostic support. This paper presents a novel approach to neurological disease classification utilizing voice recordings of individuals diagnosed with various neurological conditions alongside healthy controls. By employing AI techniques, particularly a hybrid deep network framework integrating Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), we aimed to classify one-sentence audio inputs of Multiple Sclerosis (MS) patients, healthy individuals, and other neurological diseases. In our dataset, we have compiled audio recordings from 95 healthy individuals, 99 individuals diagnosed with multiple sclerosis (MS), and 96 individuals with other neurological disorders. Of these, 20 % of the data was reserved for testing. Our proposed architecture achieved remarkable performance metrics in experimental evaluations, exhibiting 96.55 % accuracy, 98.25 % specificity, 96.49 % sensitivity, 96.97 % precision, and 96.56 % F1-Score. The results obtained are more successful compared to the methods of AlexNet from scratch, fine-tuned AlexNet, Long Short-Term Memory (LSTM) based CNN, and Gated Recurrent Unit (GRU) based CNN. The results of our study highlight the potential of this framework to be integrated into clinical diagnostic workflows, providing clinicians with an effective tool for early and precise detection of neurological diseases, ultimately improving patient outcomes through timely intervention and personalized treatment strategies.

1. Introduction

Understanding how sound is captured and processed is crucial, particularly when we consider the complexity of the human voice. Sound energy travels through a medium like air and produces changes in pressure that our ears interpret as sound (Pasnau, 1999). To store sound digitally, it must first be converted into a sequence of numerical values representing sound waves. This conversion process, known as digitization, involves using an analog-to-digital converter (ADC) and a microphone to convert the sound waves into electrical signals and digital data (Gilchrist, 1980). Computer algorithms can edit, process, and manipulate the resulting digital audio.

The implications of analyzing the human voice go beyond understanding speech mechanics; they can also be critical for health diagnostics. The human voice is a multifaceted and dynamic system, influenced by numerous factors such as the mathematical properties, the vocal tract's physiology, the speech environment's acoustics, and cognitive and motor processes controlling speech production (Asfiati et al., 2020). Various signal processing techniques, including spectral, prosodic, formant, and sound quality analysis, can be used to analyze the mathematical properties of the human voice. These analyses provide valuable insights for enhancing speech recognition systems, diagnosing

* Corresponding author.
E-mail address: emel.soylu@samsun.edu.tr (E. Soylu).

https://doi.org/10.1016/j.engappai.2025.110558
Received 26 November 2024; Received in revised form 7 February 2025; Accepted 10 March 2025
Available online 14 March 2025
0952-1976/© 2025 Elsevier Ltd. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

speech disorders, and investigating the cognitive and motor processes involved in speech production (Ren et al., 2022).

Given the complexity of voice changes, accurate diagnosis often requires a comprehensive approach. In some cases, changes in a person's voice may be indicative of an underlying health condition. For example, a hoarse or squeaky voice may be a sign of laryngeal cancer or upper respiratory tract infection (Santosh et al., 2022). Similarly, changes in pitch or volume may be an indication of neurological conditions such as Parkinson's disease (Pah et al., 2022; Ngo et al., 2022; Group et al., 1996; Chen et al., 2022) or MS (Kapur et al., 2020; Moumdjian et al., 2022; Yalachkov et al., 2019). However, changes in a person's voice can be caused by a variety of factors and may not indicate a specific disease (Asiaee et al., 2022). For example, voice changes may be the result of stress, fatigue, or environmental factors such as exposure to smoke or pollution.

In parallel to the advancements in medical diagnostics, computerized audio processing has emerged as a powerful tool in many applications. To determine the cause of the change in a person's voice, the doctor will typically do a thorough inference, including a physical checkup and, if necessary, additional testing such as imaging or blood work. Thus, while changes in a person's voice can be a useful diagnostic tool in some cases, it is not always possible to diagnose a disease based on a person's voice alone. A complete medical evaluation is typically required to make an accurate diagnosis.

Computerized audio processing has many advantages. Computers can process audio much faster and more accurately than humans, making it possible to transcribe large volumes of speech in real-time. This is particularly useful in applications such as speech recognition, where a computer can copy speech with high accuracy in a fraction of the time it takes a human to perform the same task. Computers can process sound consistently, making it possible to achieve highly repeatable results. This is important in applications where a consistent response is critical, such as speech recognition systems or speech-based emotion recognition systems. Computers can be designed to be highly resilient to environmental conditions, such as background noise in noisy environments. Computers are objective in processing sound; this means that they are not affected by personal biases or individual differences in hearing. This is particularly useful in applications where large amounts of data need to be processed, such as speech-based disease diagnosis systems (Li et al., 2020; Xiao et al., 2019; Mauch et al., 2015; Nabih-Ali et al., 2017; Gökçen, 2021; Islam et al., 2018).

Detecting a disease in its early stages can lead to more effective treatment and a better prognosis, prevent the disease from becoming more severe, and reduce the cost of treatment (Muneeb Hassan et al., 2023; Hassan et al., 2023). Early detection can save lives. In most cases, more treatment options are available in the early stages of a disease than in the later stages. Early diagnosis and treatment can help prevent complications and improve a patient's quality of life. Early detection gives patients and their families time to make informed decisions and prepare for the journey ahead. Early diagnosis can provide early intervention that can help prevent disease progression (Chu and others, 2012; Hadders-Algra, 2014; Philip and Hewitt, 1980; Becker et al., 2002; Igra et al., 2017; Brooks, 1998; Miller, 2004). Unfortunately, as a disease progresses, there may be fewer treatment options and those available may be less effective. Late diagnosis can lead to a more serious form of the disease that can be more difficult to treat and lead to more complications. Late diagnosis can reduce the chances of a full recovery and make it difficult for patients to return to their normal lives. Late diagnosis can lead to a longer recovery time as well as more extensive and costly treatments, resulting in a longer and more difficult course of treatment that can harm the patient's physical and emotional well-being, and increase the risk of death from the disease, especially for those with chronic conditions (Nzwalo et al., 2014; Cárdenas-Robledo et al., 2021; Laakso et al., 2000; Kennedy, 2013; Richards et al., 2020; Adamec et al., 2013).

Voice analysis has emerged as a valuable tool in detecting and monitoring neurological diseases. People with neurological disorders often complain of voice disturbances before other symptoms of the disease develop (Yalachkov et al., 2019; Orozco-Arroyave et al., 2015; Lopes et al., 2016). Neurological voice disorders may originate from the central nervous system, or peripheral nervous system, or be functional or behavioral (Wang and Song, 2022). Voice changes play an important role in the early diagnosis of neurological disorders. Sound tests can be used to detect neurological disease as an adjunct to other methods. Research has indicated the potential of using voice as a biomarker for various neurological disorders (Gullapalli and Mittal, 2021). Maximum performance tasks have been recognized as a domain where early signs of neurological diseases can be identified through speech and voice samples (Karlsson and Hartelius, 2021). This is supported by the fact that many neurological diseases impact phonation, making voice a potentially valuable aid in diagnosing such conditions (Rosen et al., 2005). Moreover, respiratory muscle training (RMT) has shown effectiveness in improving voice quality following neurological insults, such as dysphonia after a cerebrovascular accident (Arnold et al., 2023). Studies have also demonstrated that individuals with neurological diseases, like Parkinson's, may exhibit low-frequency vocal modulations in their speech, indicating potential differences from healthy speakers (Cnockaert et al., 2008). This emphasizes the importance of analyzing voice characteristics for early detection and monitoring of neurological disorders. Advancements in technology, such as AI and machine learning, have been utilized to develop models for classifying neurological disordered voices based on time and frequency domain features (K and Holi, 2015). Additionally, e-health tools have been developed to aid in the early identification and classification of voice pathologies and neurological diseases through speech processing (Bouafif and Ellouze, 2018).

Audio data can be captured non-invasively, making it possible to detect diseases without the need for physical examination or testing. This can be especially helpful in early diagnosis where the disease may not be easily observed. Audio data is generally less expensive to collect and process than picture data, making it a cost-effective option for disease detection. Audio data can be captured using simple equipment such as a microphone, making it more accessible to a wider range of individuals and communities. Audio data can be captured remotely, allowing disease detection without the need for physical presence of the patient. This can be especially helpful when patients are unable to visit a healthcare provider in person. AI algorithms trained on sound data can be used to detect diseases with greater accuracy, as they can analyze speech patterns, rhythm, and other acoustic cues that may indicate the presence of a disease. Overall, while image data has its advantages in disease detection, the benefits of robust data such as non-invasiveness, cost-effectiveness, and remote detection make it a valuable resource in healthcare and medicine.

Sound classification automatically sorts sounds such as speech, music, environmental sounds, etc. Many methods are used in sound classification. The feature extraction method involves extracting several handcrafted features from the audio signal, such as Mel-frequency cepstral coefficients (MFCCs), spectral center, spectral flatness, and zero-crossing ratio. These features are then used as input to a machine-learning algorithm for classification (Bezoui et al., 2016; Deng et al., 2020; Nagawade and Ratnaparkhe, 2017; Shi et al., 2018; Prabakaran and Sriuppili, 2021). Deep learning techniques such as CNNs and RNNs show great success in audio classification tasks. In the deep learning technique, feature extraction is done automatically (Khamparia et al., 2019; Zhang et al., 2017a; Piczak, 2015; Huzaifah, 2017; Su et al., 2019). Some sound classification methods are hybrid, both feature extraction methods and deep learning methods can be used together. In such approaches, raw data and features constitute the network input (Sharma et al., 2020; Tokozume et al., 2017; Purwins et al., 2019). Another method, the transfer-based learning method, involves using a deep learning model on a model that was previously trained with a large dataset and then fine-tuning the model on a smaller, task-specific

dataset. Choosing the appropriate one among these sample techniques depends on the specific problem and the type of data being analyzed (Zhuang et al., 2020; Liang and Zheng, 2020; Lee et al., 2018; Zhang et al., 2017b).

Machine learning and deep learning methods are applicable across various domains and have demonstrated successful outcomes (Akkem et al., 2023a, 2023b, 2024). AI has the potential to revolutionize healthcare by providing new and improved methods for disease diagnosis, treatment, and prevention. AI can analyze large amounts of medical data, such as imaging studies and electronic health records, to help identify disease patterns and symptoms. This could lead to more accurate and earlier diagnoses. Advanced CNN architectures can support early diagnosis of various diseases. Brain tumor classification and early diagnosis of chronic kidney diseases can be cited as examples of these studies (Mahmoud et al., 2024; Sheela et al., 2024). AI can analyze a patient's medical history, genetics, and other personal data to develop individualized treatment plans and predict potential health risks. AI analyze clinical trial data more efficiently and accurately, enabling faster and more cost-effective drug development. AI can be used to develop predictive models that help healthcare providers make more informed decisions about patient care and improve patient outcomes. AI can be used to provide real-time support to healthcare providers by analyzing patient data and providing relevant information and recommendations. Neurological disease detection through voice analysis using AI has gained significant attention in recent research. AI algorithms have been applied to voice data to detect various neurological conditions, showcasing the potential for innovative diagnostic tools in the field of neurology (Verde et al., 2018). Voice analysis has been utilized in the identification of neurological disorders, including dysphonia, through machine learning techniques, demonstrating the feasibility of using AI for voice disorder detection (Verde et al., 2018; Chen et al., 2023). Some studies demonstrate the effectiveness of machine learning models like neural networks and Gaussian process regression in accurately predicting complex patterns across various domains, highlighting their potential applicability in neurological research, particularly for modeling and understanding sound processing in the brain (Jin and Xu, 2024a, 2024b, 2024c, 2024d, 2024e, 2024f, 2024g; Xu and Zhang, 2021).

In addition to the techniques mentioned above, hybrid deep learning methods are very successful in sound classification. They combine the strengths of different models to effectively capture both the spatial and time-related features of audio signals. The advantages of hybrid CNN, RNN, and LSTM models in sound classification stem from the synergistic effects resulting from the combination of their individual characteristics. These hybrid approaches enable more effective processing of the complex structure of sound data and improve classification performance. Firstly, CNNs have the ability to process the time-series characteristics of sound data in a visual manner. This is particularly important when extracting the spectral features of sound waves. CNNs are highly effective in recognizing local features, which are crucial for analyzing sound signals (Bilgera et al., 2018; Prasad and Deepa, 2021). For example, a CNN-LSTM hybrid model extracts features in both the time and frequency domains of sound signals, resulting in improved classification performance. The ability of CNNs to capture spatial relationships in sound data, combined with the ability of LSTMs to model temporal dependencies, provides a significant advantage in sound classification (Prasad and Deepa, 2021; Kim et al., 2018). LSTM, as an extension of RNN, is known for its ability to learn long-term dependencies. This capability is crucial for understanding the temporal variations in sound signals. LSTM is particularly successful in modeling complex temporal dependencies in sound signals (Süzen et al., 2019). For instance, LSTM-based models achieve higher accuracy rates in speech recognition and classification tasks compared to traditional RNNs (Prasad and Deepa, 2021; Wang et al., 2020). Additionally, hybrid CNN-LSTM models enhance classification accuracy by simultaneously processing the spatial and temporal features of sound data (Bilgera et al., 2018; Nguyen-Da et al., 2023). Another advantage of hybrid models is that they combine the strengths of different deep learning architectures, providing a more comprehensive learning process. For example, the CNN-LSTM combination effectively learns both local and global features of sound data, significantly improving classification performance (Bilgera et al., 2018; Prasad and Deepa, 2021; Nguyen-Da et al., 2023). Furthermore, these hybrid structures enhance the model's generalization capability, enabling better results with less data.

The motivation for this study is to explore the potential of voice analysis and deep learning techniques in the early diagnosis of neurological diseases. If neurological diseases can be detected in their early stages, the treatment process can be more effective and efficient. Traditional diagnostic methods can often be invasive and costly, making the development of non-invasive and cost-effective methods like voice analysis highly advantageous for both patients and healthcare systems. In this study, deep learning models with high accuracy rates have been developed to analyze voice data for the detection of various neurological diseases. This approach not only increases the accuracy and reliability of the diagnosis process but also offers innovative methods that can be effectively used in clinical applications. Moreover, the results of this study demonstrate the future potential of voice analysis and AI-based approaches in the diagnosis of neurological diseases, contributing significantly to the existing literature in this field.

In the literature, there are many studies on neurological disease detection using different methods. Among these studies, the high success rates of machine learning and deep learning techniques draw attention. Obtaining helpful comments on the disease with voice analysis is a painless and cost-effective method for the patient. In this study, we propose a deep learning model with a higher accuracy rate than other studies, with the use of an original data set. The data set used in the study was obtained from more subjects than many studies in the literature. Since the number of individuals with MS disease among individuals diagnosed with neurological disease is high, this disease has been evaluated as a separate class and belongs to other diseases such as ALS, SCA, Alzheimer's disease, Epilepsy, Parkinson's, Myasthenia Graves, Myelitis, Motor Aphasia, Psychological, Friedreich Ataxia, Language Problem. Since the number of subjects was small, they were evaluated as a single group. In addition, a healthy class was created by taking voice recordings from healthy subjects.

The main contributions of this study are:

- This study uses CNN and RNN-based deep networks for the accurate detection of nervous system diseases through audio signals. The high accuracy and precision values obtained as a result of the experimental evaluations emphasized the effectiveness and reliability of this approach.
- The proposed model offers a hybrid network approach by combining different deep learning architectures such as CNN and RNN. This allows for more effective analysis and classification of audio signals in the process of detecting nervous system diseases.
- The extension of the dataset to include multiple neurological disorders and the definition of specific disease classes significantly enhance the model's diagnostic capabilities. This contribution is crucial as it improves the model's accuracy and generalizability, making it a more effective tool for clinical applications. The expanded dataset enables the model to better characterize and distinguish between a wider range of diseases, thus providing a robust framework for future research and practical use in healthcare.
- The study provides a comprehensive analysis of how AI technologies, specifically voice analysis, can be effectively utilized for diagnosing neurological diseases. By achieving high performance metrics, the research demonstrates the potential for these technologies to be incorporated into clinical settings, thus paving the way for practical applications.
- The research contributes to the development of advanced diagnostic technologies by showcasing the capabilities of AI in early disease detection. The study not only validates the effectiveness of AI-based

approaches but also sets a foundation for future advancements in diagnostic methods, emphasizing their importance in improving healthcare practices.

In the following sections of the article, there are acoustic characteristics of healthy speech and neurological disorder-related speech, related work, background, proposed model, experimental findings, and conclusion sections. In the background section, the techniques used in this study, which constitute the infrastructure of this study, are mentioned. In Section 4, the proposed model is detailed step by step. In the experimental studies section, the dataset, the classification process, and the results obtained are given. Finally, the result of the study, its contribution to the literature, and how it can be used for further studies are interpreted.

2. Acoustic characteristics of healthy speech and neurological disorder-related speech

Acoustic markers provide a valuable and non-invasive approach to detecting, diagnosing, and monitoring neurological diseases (Kent and Rosenbek, 1982). Their theoretical basis lies in the disruption of neural circuits responsible for motor control, executive function, and language processing. Clinically, acoustic biomarkers hold promise for early disease detection, tracking disease progression, and evaluating treatment effectiveness.

Neurological speech disorders exhibit a range of distinctive acoustic characteristics, reflecting the underlying neural impairments. Speech rate can be slow and irregular, as seen in conditions like multiple sclerosis (MS) (Renauld et al., 2016), spinocerebellar ataxia (SCA), and Friedreich's ataxia, or it may be accelerated, such as in Parkinson's disease (Rusz et al., 2011), where speech often occurs in short rushes. Pitch variation is generally reduced in disorders like ALS (Norel et al., 2018), MS, Parkinson's disease (Majda-Zdancewicz et al., 2024), and myelitis, leading to a monotonous and flat vocal tone. However, excessive pitch variation is observed in ataxic disorders like SCA and Friedreich's ataxia (Ball et al., 2002; van Prooije et al., 2024), resulting in unpredictable and scanning prosody.

Articulation deficits are a common feature across various neurological conditions, with imprecise or slurred speech appearing in ALS, myasthenia gravis, MS, and ataxic disorders. These deficits often lead to reduced speech intelligibility, particularly in conditions affecting motor control and coordination. Voice quality varies significantly depending on the disorder, ranging from breathy and weak phonation in ALS and myasthenia gravis to harsh and strained vocal production in MS, SCA, and Friedreich's ataxia. Additionally, dysphonia, hypernasality, and vocal fatigue are frequently present, particularly in progressive neuromuscular diseases.

Prosody and intonation are often disrupted, with monotonous speech commonly observed in ALS, Parkinson's disease, and MS, while scanning speech, characterized by equal stress distribution, is typical of cerebellar disorders. Some conditions, such as Alzheimer's disease and epilepsy, also present with impaired speech planning and increased pauses. Psychological and language-related disorders, while not necessarily affecting neuromuscular function, may still lead to exaggerated or unstable prosody, speech disfluencies, or difficulty with word retrieval.

By contrast, healthy individuals demonstrate fluent articulation with clear and precise pronunciation of consonants and vowels. Their speech rate remains balanced, ensuring smooth transitions between syllables, words, and sentences. Adequate voice control allows for consistent loudness and pitch modulation, while stable resonance prevents excessive nasality or breathiness. Moreover, strong phonatory control ensures steady phonation, preventing voice tremors or harshness, resulting in natural and effortless communication (Duffy and others, 2012).

These acoustic characteristics reflect proper motor coordination, respiratory function, and intact neurological control, highlighting the fundamental differences between healthy and pathological speech. As illustrated in Table 1, the comparative analysis of acoustic features in neurological speech disorders highlights the distinct variations in speech rate, pitch variation, articulation, voice quality, prosody, and other acoustic characteristics across different conditions (Kent and Rosenbek, 1982; Renauld et al., 2016; Rusz et al., 2011; Norel et al., 2018; Majda-Zdancewicz et al., 2024; Ball et al., 2002; van Prooije et al., 2024; Duffy and others, 2012; Goberman and Coelho, 2002; Lehman et al., 2020; Freed, 2023; Brabenec et al., 2017).

3. Related work

In recent years, deep learning and machine learning techniques have played a significant role in the classification of neurological disorders using audio data. In this context, systematic reviews of Alzheimer's disease and mild cognitive impairments have shown that automated audio analysis methods are effective in diagnosing these conditions (Faust et al., 2018; Hong et al., 2019). Specifically, deep learning algorithms have yielded promising results for the early diagnosis of Alzheimer's disease through the analysis of audio data. For example, models for diagnosing Alzheimer's disease have been developed using deep learning techniques such as LSTM (Hong et al., 2019). Similarly, studies on the classification of Parkinson's disease using audio data demonstrate the effectiveness of deep learning techniques in this area. For instance, high accuracy rates have been achieved in diagnosing Parkinson's disease through the analysis of audio data (Rahman et al., 2023; Syed et al., 2020). Additionally, identifying different stages of Parkinson's disease through audio analysis is crucial for the management and personalization of treatment processes (Maiti, 2024; Mahmood et al., 2023).

MS is a neurological disease that typically occurs in young adults. It is characterized by demyelination and increasing axonal degeneration in the central nervous system. The disease is considered epigenetic and is caused by the influence of environmental factors in individuals with genetic susceptibility. It is a major cause of disability in young people and is more prevalent in women than men. The disease can affect multiple areas of the central nervous system, leading to speech and voice problems that vary depending on the affected anatomical region. A single diagnostic test for MS does not exist, and the diagnosis is based on factors such as medical history, examination, and MRI of the brain and spine, as well as cerebrospinal fluid analysis. Given the heterogeneous distribution and clinical course of the disease, it is important to establish a personalized treatment plan for patients by determining their prognostic process. Early implementation of treatments such as pharmacological and neurorehabilitation interventions can have a positive impact on the disease's progression.

Swallowing, speech, and voice problems are common in MS patients, negatively affecting their quality of life. Early analysis of speech and voice problems is therefore critical. Speech analysis can help diagnose the disease correctly and determine the extent of impairment, thereby enabling the addition of effective neurorehabilitation practices to the treatment process in the early stages. Vocal problems are also common in MS patients and may result in language skills being affected. Accordingly, in the current study, an AI based method was developed to assist in the diagnosis of individuals through voice recordings. Additionally, the literature studies conducted in this field have been summarized in Table 1, and their performance results have been compared with those of the proposed study.

Table 2 shows examples of studies using audio data in the classification of neurological disorders. These studies include Alzheimer's disease, Parkinson's, MS, Amyotrophic Lateral Sclerosis, and other neurological diseases. Statistical techniques, machine learning, and AI techniques were used in sound classification. It is seen that the success rates are high in studies where deep learning techniques are used.

Our study makes several unique contributions to the field of neurological disorder classification using voice analysis. In comparison to the studies in the table, our research used a unique dataset, which allowed

Table 1
Comparative table of acoustic features in neurological speech disorders.

Category	Speech Rate	Pitch Variation	Articulation	Voice Quality	Prosody & Intonation	Other Acoustic Features
Normal Speech	Normal	Normal	Clear	No dysphonia	Normal	No abnormalities
Multiple Sclerosis (MS)	Slow, irregular	Reduced variation	Slurred speech, dysarthria	Harsh, strained voice	Monotonic, flat	Reduced speech intelligibility
Amyotrophic Lateral Sclerosis (ALS)	Gradual slowing	Reduced variation	Imprecise articulation	Weak, breathy voice	Monotone, loss of prosodic features	Hypernasality, decreased loudness
Parkinson's Disease	Accelerated (short rushes)	Reduced variation	Imprecise consonants	Breathy, weak phonation	Flat, monotonous	Hypophonia, reduced volume
Alzheimer's Disease	Normal/slightly slow	Reduced variation	Word-finding difficulties	Normal/mild dysphonia	Impaired speech planning	Increased pauses, anomia
Epilepsy	Normal/slightly slow	Unstable	Normal/slightly impaired	Variable	Disorganized in ictal state	Sudden speech arrest
Myasthenia Gravis	Fatigue-related slowing	Normal/ reduced	Slurred, weak articulation	Breathy, fatigued voice	Reduced inflection	Worsens with prolonged speech
Spinocerebellar Ataxia (SCA)	Slow, irregular	Excessive variation	Distorted articulation	Harsh, strained	Scanning speech (equal stress)	Uncoordinated voice breaks
Friedreich's Ataxia	Slow, irregular	Excessive variation	Imprecise, jerky articulation	Harsh, strained	Scanning, unpredictable prosody	Dysdiadochokinesia
Myelitis	Variable	Reduced variation	Impaired articulation	Strained or breathy	Flat	Can resemble spastic dysarthria
Motor Aphasia	Normal/slightly slow	Normal	Difficulty in word retrieval	Normal	Simplified sentence structure	Difficulty initiating speech
Psychological Disorders	Variable	Normal/ exaggerated	Normal/slightly impaired	Varied, sometimes tremulous	Unstable, exaggerated or flat	Influenced by emotional state
Language Disorders	Variable	Normal	Impaired phoneme production	Normal/slightly hoarse	Affected by linguistic deficits	Speech disfluency, mispronunciations

us to test different methods. While previous studies worked with established datasets like the Saarbruecken Voice Database or Dem@-Care, we created and used our own data. This makes our approach different and required us to compare various techniques to see how well they perform on new data. As a result, our study adds a new perspective by evaluating how different models handle a unique dataset, contributing valuable insights to the classification of neurological diseases.

In our study, unlike these studies, the classification of MS, healthy individuals, and other neurological disorders is done at the same time and the number of sample individuals is above the average. With the CNN-based proposed architecture, in which the GRU and LSTM models we used in our study, a success rate of 96.6 % was achieved.

We conducted the first study evaluating the potential contribution of voice analysis in differential diagnosis by including more MS patients and evaluating patient groups with different neurological disorders and healthy people. Unlike other studies, we aimed to distinguish MS patients from those with different neurological disorders with voice abnormalities, achieving high success rates. With larger patient data in the future, the contribution of voice analysis to the diagnosis and prognostic process of neurological disorders can be demonstrated more clearly using machine learning and AI based methods. By teaching the machine with patients' demographic, clinical, walking, affected neurological system, radiological, and laboratory data as well as voice analysis, it will provide a significant contribution to determining individual prognostic processes within the complex heterogeneity of the disease. These data will be useful in both differentiating diseases and monitoring the clinical course of existing diseases.

4. Background

The theoretical background and materials of the proposed model in this study are presented in subtitles. In the study, Short-time Fourier transform (STFT) was used to convert audio data to image. CNN and RNN were used to classify images.

4.1. Short-Time Fourier Transform

The Fourier transform allows the signals in the time domain to be expressed in the frequency domain. It was invented by the French mathematician and physicist Jean Baptiste Joseph Fourier. The Fourier Transform makes it possible to split the original time signal into sinusoids. Each sinusoid has an associated amplitude, phase, and frequency. A complex-looking waveform in the time domain can be represented by a vertical line in the frequency domain. This simple representation in the frequency domain helps identify key frequencies. The Fourier Transform separates complex time signals into frequency components, making it easy to understand. No data is lost when moving from the frequency domain to the time domain.

Audio signals are variable, not static. So, the stats of the sound change over time. It would be pointless to calculate a single Fourier transform over the entire speech in a 5-min lecture. Such a transformation will not provide distinguishable data in the data analysis process. Instead, the Fourier transform of consecutive frames in the signal is performed. This method is called STFT (Stft. https).

$$X(m,\omega) = \sum_n X(n) w(n-m) e^{-j\omega n} \tag{1}$$

Eq (1) breaks down a signal into its time and frequency components. In this equation $X(m, \omega)$ is the output of the STFT, where m represents time frame and ω represents frequency, n is sampling index, $X(n)$ is the nth sample of the input signal, $w(n\text{-}m)$ is the window function, used to select a portion of the signal, j is imaginary unit and ω is frequency. As we increase m in Eq (1), we shift the frame function w to the right. We calculate the Fourier transform for the resulting frame $x(n)\ w(n-m)$. Therefore, STFT X is a function of both time (m) and frequency (ω).

In our study, the STFT process was performed using the librosa library. librosa.stft calculates a STFT. For calculation, the function is provided with a frame size, i.e. the size of the FFT, and a hop length, i.e. frame increment.

$$S(m,\omega) = |X(m,\omega)|^2 \tag{2}$$

When processing sound, the phase content is often disregarded in favor of only the spectral size. A spectrogram is a visual representation of the intensity of frequencies over time and is derived from the squared size of the STFT, as described in Eq (2). This equation represents the process of generating a spectrogram by taking the square of the absolute value of the STFT. $S(m, \omega)$ is the value of the spectrogram. This tool combines the advantages of both the time and frequency domains and can display changes in the speech spectrum over time. Spectrograms are useful for segmenting information from audio files and converting them

Table 2
Comparison of previous studies of the proposed model.

Classes	Data size	Technique	Acc. (%)	Ref.
neurological diseases	Saarbruecken Voice Database and vocal sound recorded from over 2000 healthy and pathological subjects	Deep Learning	82.69	Alghamdi et al. (2022)
Amyotrofik Lateral Skleroz	100 utterances each from male and female speakers	Machine Learning	8 for females and 79 % for males	Norel et al. (2018)
Amyotrofik Lateral Skleroz	syllable repetition paradigm collected from 18 patients	Deep Learning	87.6	Novotny et al. (2020)
Alzheimer's disease	VBSD dataset (504 speech data), Dem@Care dataset (32 speech data)	Machine Learning	86.1	Liu et al. (2020)
MS	85 MS patient	Statistical Methods	84.9	Noffs et al. (2020)
Alzheimer's disease	43 patients	Machine Learning	81.8	Frid et al. (2014)
Parkinson	64 Parkinson's disease patients, 34 HC – healthy control subject	Machine Learning	94.55	Almeida et al. (2019)
Parkinson	23 people with Parkinson's disease, 8 health controls	Machine Learning, Deep Neural Network	98	Haq et al. (2018)
Parkinson	50 Parkinson's disease patients, 50HC – healthy control subject	Transfer Based Deep Learning	99	Zahid et al. (2020)
MS	22 MS subjects, 19 healthy controls	Deep Neural Network	92.6	Gosztolya et al. (2022)
MS	30 MS individuals, 76 healthy individual	Statistical Methods	70	Feijó et al. (2004)
MS	65 MS individuals, 66 healthy individual	AI Techniques	82	Yamamoto et al. (2010)
MS, Parkinson and healthy	95 healthy individuals, 99 MS individuals, and 96 neurological disease individuals	RNN-based CNN model	96.55	**Our**

into images. They typically have a horizontal axis representing time, a vertical axis representing frequency, and color intensity representing the amplitude of a frequency at a given point in time. The spectrogram displays the energy distribution of different frequency signals, and the intensity of any frequency component is expressed through color depth. Soundtracks, which are represented by different lines, can be created by varying shades of black and white within the spectrogram (Ye and Yang, 2021; Li, 2011).

4.2. Convolutional Neural Networks

Artificial neural networks (ANNs) are designed to emulate the functioning of biological nervous systems. ANNs are composed of interconnected computational nodes or neurons that work together in a distributed manner to learn from inputs and optimize their outputs. Similarly, CNNs update their weights through learning and involve a series of nonlinear operations with scalar products and activation functions. However, CNNs are primarily used in image recognition tasks (O'Shea and Nash, 2015). Tensors are of the utmost importance within CNNs. Tensors with dimensions exceeding 3 are prevalent in CNNs, and tensors encompass the input data, intermediate representations, and parameters within a CNN. Conventionally, a CNN accepts a 3rd order tensor, exemplified by an image with dimensions H rows, W columns, and 3 channels (representing the R, G, and B color channels). Subsequently, the input undergoes a sequence of operations, where each operation is denoted as a layer. These layers encompass possibilities such as convolution layers, pooling layers, fully connected layers, loss layers, and more.

In a convolutional layer, a filter (also known as a kernel) slides over the input image or feature map, and a dot product is computed at each position. This process helps in detecting various features in the input. Mathematically, the operation at position (i, j) in the feature map can be represented as Eq (3). (Şakar HG Evrişimli Sinir Ağları):

$$(F \bullet I)_{ij} = \sum_{m} \sum_{n} F_{mn} \bullet Ii + m, j + n \tag{3}$$

where $(F \cdot I)_{ij}$ represents the value at position (i,j), in the resulting feature map. F_{mn} is the filter's value at position (m,n), $I_{i+m,j+n}$ is the input value at position $i + m, j + n)$. Typically, after the convolution operation, an activation function is applied elementwise to introduce non-linearity into the network. One common activation function is the Rectified Linear Unit (ReLU) as given in Eq (4):

$$ReLU(x) = \max(0,x) \tag{4}$$

Pooling layers reduce the spatial dimensions of the feature maps, helping to decrease the computational complexity and control overfitting. Max pooling is a common pooling operation as Eq (5):

$$MaxPooling(x) = \max(x) \tag{5}$$

Fully connected layers are traditional neural network layers where each neuron is connected to every neuron in the previous layer. If we have a flattened vector of features x from the previous layers, the output of a fully connected layer can be represented as Eq (6):

$$Y = Wx + b \tag{6}$$

Where Y is the output vector, W is the weight matrix, x is the input vector and b is the bias vector. The loss layer computes the difference between the predicted output and the actual target labels. Common loss functions include Mean Squared Error (MSE) for regression tasks and Cross-Entropy Loss for classification tasks. For example, MSE can be computed as given in Eq (7):

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \tag{7}$$

where N is the number of samples, y_i is the actual target for sample I, $\hat{y}_i$ is the predicted output for sample i. These equations encompass the core operations of a CNN. CNN architectures may integrate modifications and extra layers, but these remain the fundamental equations.

4.3. Recurrent neural network (RNN)

Artificial neural networks are trained in 3 basic steps. In the first step forward, action is taken to make predictions. Secondly, using the (Loss) function, this estimate is compared with its true value. After this, the loss function returns an error value. Finally, using this error value, the gradient is calculated for each node in the network, and this happens as backpropagation. In contrast, an RNN contains a latent state that feeds information from previous states. This concept of latent state is like integrating sequential data to make a more accurate prediction. If we have sequence information, the predictions will be more accurate. (RNNs are widely used in research areas that analyze sequential data such as text, audio, and video.

CNNs characterized by having acyclic connections between their modules or layers. This means that when computing the outputs, there is a partial order that ensures that the inputs are available when needed. In contrast, RNNs allow for loops in their connections, as shown in Fig. 1.

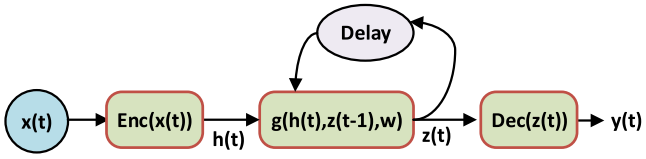

Fig. 1. An RNN with a loop.

The equations of *zt, rt, ht* are given in Eq (8), Eq(9), and Eq (10) respectively. In an RNN, the input *x(t)* varies over time, and the encoder *Enc(x(t))* produces a representation *h(t)* of the input using trainable parameters *w*. The function that computes the current hidden state *z(t)* based on the previous hidden state *z(t-1)* and the current input *h(t)* can be a complex neural network, with one of its inputs being the previous time step. Finally, the decoder *Dec(z(t))* generates the output. The architecture of RNNs causes an unavoidable issue called the vanishing gradient problem during back-propagation. As the weights in each layer are updated using the Chain Rule, the gradient values tend to decrease exponentially as the network goes backward, causing them to "disappear" eventually. One of the benefits of using RNNs is the ability to retain information from the past. However, if the gradients are not truncated or initialized correctly, the RNNs may not be able to remember information from the distant past. RNNs suffer from problems like vanishing/exploding gradients and poor ability to retain long-term states. To address these issues GRU, which is an application of multiplicative models, has been developed (Cho et al., 2014). GRU is a type of recurrent network that incorporates memory. The structure of a GRU unit can be observed in Fig. 2.

The architecture of RNNs leads to the issue of disappearing weights during back-propagation. This occurs because the gradient values decrease exponentially as we move backwards in the network and can eventually vanish, preventing the network from remembering information from the past. To address this issue, multiplicative models such as GRU have been developed. GRU is a type of recursive network that incorporates memory and helps to overcome the problems of vanishing/exploding gradients and the inability to remember long-term states in RNNs (Cho et al., 2014). The structure of a GRU unit can be seen in Fig. 2, where $\odot$ denotes the Hadamard product, x_t represents the input vector, ht is the output vector, z_t is the update gate vector, r_t is the reset gate vector, ϕ_h is hyperbolic tanh, and W, U, b are the learnable parameters. The update gate vector z_t determines how much of the past information should be transferred to the future. It applies the sigmoid function to the sum of two linear layers: one with a bias to the x_t input and the other with the past ht-1 state. The resulting z_t coefficients range between 0 and 1, where a value of 1 indicates a copy of the previous volume and a value less than 1 indicates that some of the input information should be considered. The reset gate r_t determines how much of the past information should be forgotten. If the coefficient ϕ_h $(w_h xt + U_h(r_t \odot h_{t-1}) + b_h)$ is 0, the information from the past is not used in the new memory content. If z_t = a, h_t only looks at the input, and the whole system is reset.

$$zt = \sigma_g(W_z x_t + U_z h_{t-1} + b_z) \tag{8}$$

$$rt = \sigma_g(W_r x_t + U_r h_{t-1} + b_r) \tag{9}$$

$$h_t = z_t \odot h_{t-1} + (1 - z_t) \odot \varphi_h(W_h x_t + U_h(r_t \odot h_{t-1}) + b_h) \tag{10}$$

LSTM is another type of model that addresses the long-term memory loss issue in RNNs by creating memory cells that store past information. The LSTM structure is depicted in Fig. 3 and is more complex than GRU. Despite its later development, GRU is a simplified version of LSTM (Hochreiter and Schmidhuber, 1997). The capability of RNNs that use sigma or tanh cells to learn significant information from input data diminishes when there is a large gap between inputs. However, the inclusion of gate functions within the cell architecture has allowed for LSTM to effectively handle long-term dependencies. LSTM has emerged as the go-to approach for achieving successful results in RNN applications, with almost all successful applications using this technique since its inception (Yu et al., 2019).

Eq 11–15 defines the Hadamard product represented by the symbol $\odot$. In the LSTM unit, x_t denotes one of the inputs, f_t is the activation function of the forget gate, it is the activation vector of the input update gate, o_t is the activation vector of the output gate, h_t is the hidden state vector, and ct is the state vector of the cell. To transmit information within the LSTM unit, the cell status c_t is utilized, which is regulated by gates that determine whether the information is to be preserved in the cell state. The forget gate f_t decides the information to keep from the previous cell state c_{t-1} based on the current input and the previous hidden state and expresses it with a value between 0 and 1 with the coefficient c_{t-1}. A candidate for updating the cell state is calculated using tanh $(W_c xt + U_c h_t\text{-}1 + b_c)$, and the input gate decides how many updates to apply, similar to the forget gate. Finally, the output h_t is computed based on the cell state c_t, passed through tanh, and filtered by o_t.

$$f_t = \sigma_g(W_f x_t + U_f h_{t-1} + b_f) \tag{11}$$

$$i_t = \sigma_g(W_i x_t + U_i h_{t-1} + b_i) \tag{12}$$

$$o_t = \sigma_o(W_o x_t + U_o h_{t-1} + b_o) \tag{13}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot tan\, h(W_c x_t + U_c h_{t-1} + b_c) \tag{14}$$

$$h_t = o_t \odot tan\, h(c_t) \tag{15}$$

5. Proposed methodology

In the current study, we presented a deep approach based on a combination of CNN and RNN models for the recognition of neurological disease. The CNN layers utilize the image-like structure of the

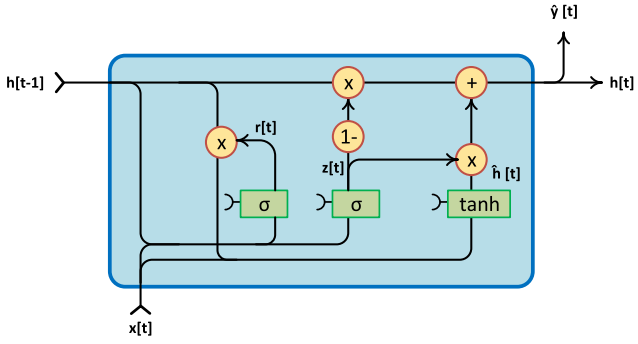

Fig. 2. Structure of the GRU architecture.

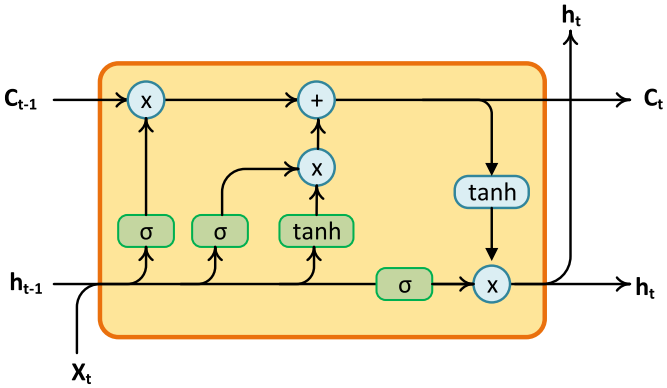

Fig. 3. Structure of the LSTM architecture.

spectrogram to extract local features of the frequency components. This helps in learning the relationships between frequency bands and the energy distribution within specific frequency ranges. The feature maps obtained from the CNN layers are fed into the RNN layers. The RNN learns the temporal changes in the spectrogram and captures temporal dependencies by utilizing information from successive time steps. The combination of CNN and RNN enables efficient learning of both spatial and temporal information, leading to higher accuracy. It is well-suited for processing complex data like spectrograms, as it can capture fine details between frequency components and temporal variations. The general representation of the proposed approach is shown in Fig. 4, and further details can be found in the sub-sections.

Step 1: *Converting audio signals to video;* The data set obtained in the form of audio recordings was converted into images before applying to the classification model. At this stage, spectrograms were obtained from each audio file by using the STFT. Sample spectrogram images for 3 classes are given in Fig. 5. Librosa library is used in this stage. The sampling rate of the audio file was used in its original form. The window size for the Short-Time Fourier Transform (STFT) was set to 2048, and the hop length, which determines how many samples away the start of the next time window is from the previous one, was set to 512.

Step 2: *CNN;* In this study, learned weights of the pre-trained deep model were used to extract significant and high-level features. Accordingly, the trained weights of the first 5 conv layers of the pre-trained model were frozen, and these layers were given to the input of the RNN-based approach. A sample illustration of the fine-tuning approach is given in Fig. 6.

Step 3: *RNN-based deep classification network;* At this stage, the 5-layer conv architecture is given to the inputs of the LSTM and GRU models in parallel. Then, a dropout layer was added after each model, and added on elementwise of 2 inputs. This addition layer was fed with 3 fully connected layers, and finally, the architecture was completed using the SoftMax layer. The general structure of the proposed architecture is given in Table 3.

In Table 2, the layers, filters, types, activations, and learnable properties used in the proposed architecture are given. This architecture contains 29 layers and 22.8 learnable.

6. Experimental works

In this paper, we presented a deep network based on CNN and RNN for audio analysis. Experimental works performed using MATLAB 2022a. In addition, a workstation computer with 32 GB RAM and NVIDIA Quadro P4000 card is used for all applications. The dataset used in this study, the numeric and visual results of the experimental studies, and performance metrics are given in sub-titles.

6.1. Dataset

Data is crucial in deep learning and is the key factor driving the performance of deep learning algorithms. Deep learning algorithms rely on large amounts of data to train models that can recognize patterns and make predictions. The more data an algorithm has access to, the more accurate its predictions can be. The data set used in this study was obtained by having healthy individuals diagnosed with MS and other neurological diseases (ALS, SCA, Alzheimer's disease, Epilepsy, Parkinson's, Myasthenia Gravis, Myelitis, Motor Aphasia, Psychological, Friedreich Ataxia, Language Problem) say a common sentence in Turkish. 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. There are.wav audio files of 95 healthy individuals, 99 individuals diagnosed with MS, and 96 individuals with other neurological diseases in the dataset. The study group consisted of healthy individuals between the ages of 18–65 who were diagnosed with a neurological disease. Table 4 presents the distribution of individuals across different groups, categorized by age range and gender. The dataset was designed to ensure balanced representation across gender and age range while maintaining the reliability and generalizability of the findings. This dataset provides a comprehensive foundation for investigating relevant characteristics across the three groups. Conversation mode includes the pronunciation of a Turkish sentence by a native speaker, "Today is very nice, tomorrow it may rain". The database contains data from a total of 290 people, both male and female, for smartphone (SP) recordings. To improve usability, audio recordings were captured using mobile phones under the supervision of expert healthcare professionals. In Fig. 8, data collection and preprocessing steps are visualized in 7 stages. While the person is saying the sentence that has been memorized, it is recorded with a smartphone. The audio recording is sent to the computer. Sound recordings are converted to. wav files on the computer. These files are stored in the corresponding folder on the computer. Each sound file has been converted into an image of 218x208 pixels, as seen in Fig. 7.

During the training, 80 % of the spectrogram images are randomly

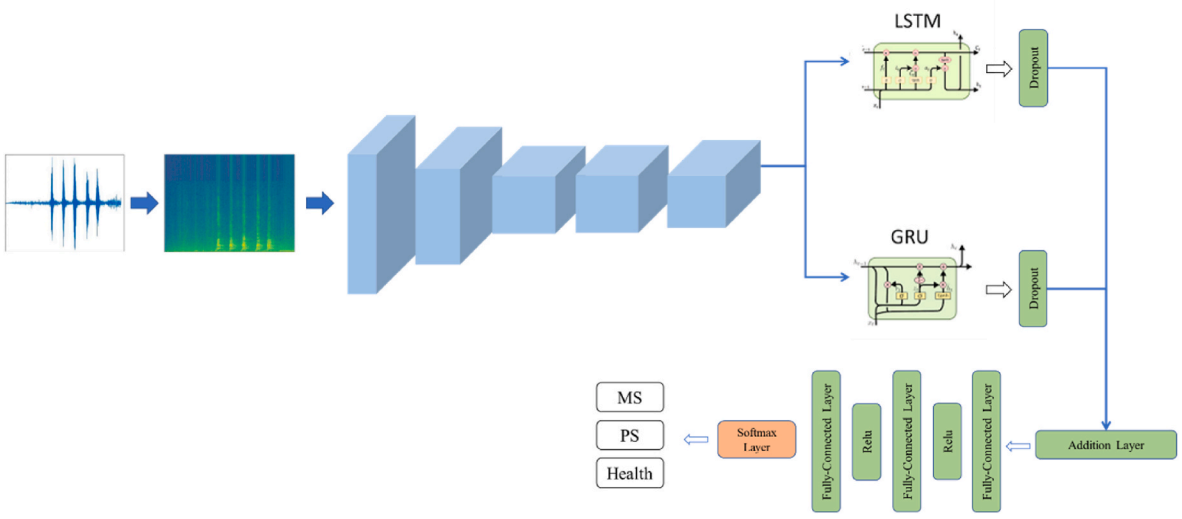

Fig. 4. The proposed architecture.

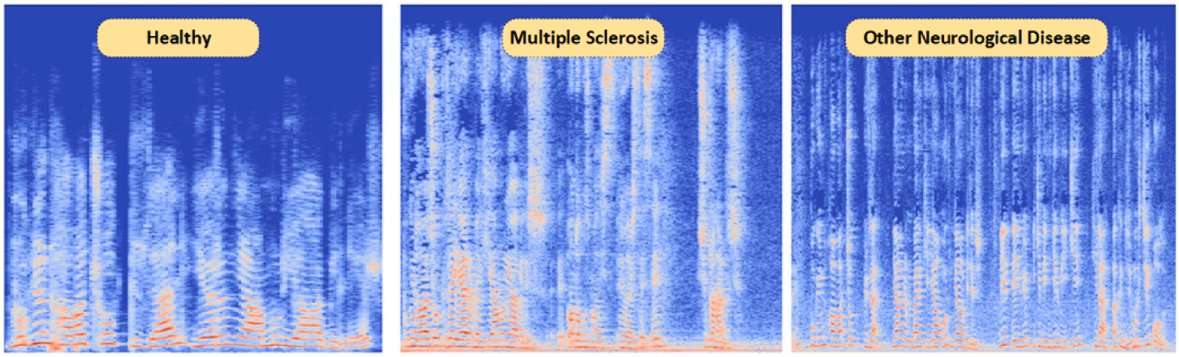

Fig. 5. Sample spectrogram images for 3 classes.

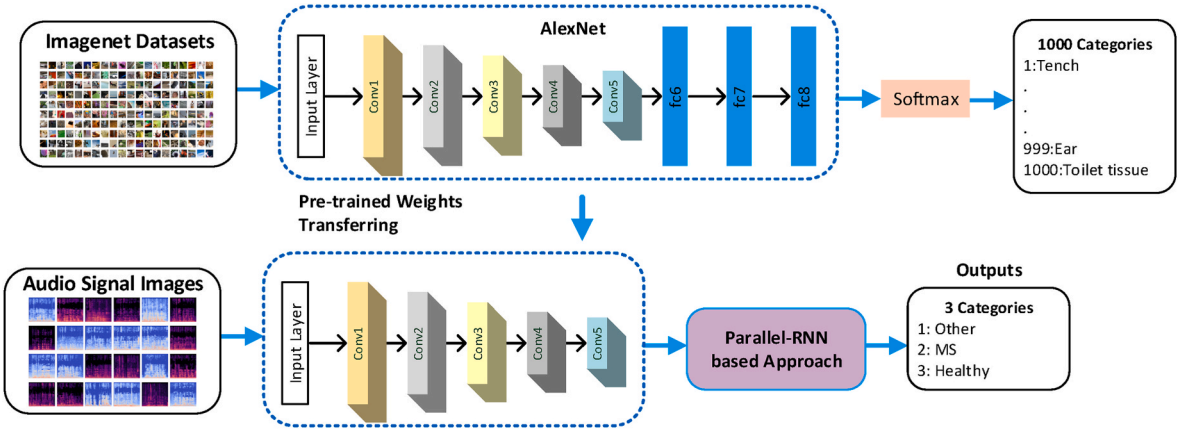

Fig. 6. Proposed model based on fine-tuning approach.

allocated as training data and 20 % as test data.

6.2. Performance metrics

Experimental works used the performance metrics that are commonly used to evaluate the performance of classification models. These:

- Sensitivity or Recall: Sensitivity is the proportion of actual positive instances that are correctly identified by the model. It is calculated as TP/(TP + FN).
- Specificity: Specificity is the proportion of actual negative instances that are correctly identified by the model. It is calculated as TN/(TN + FP).
- Precision: Precision is the proportion of instances that the model predicted as positive that are actually positive. It is calculated as TP/(TP + FP).
- F1-Score: F1-Score is the harmonic mean of precision and recall. It is calculated as 2 * (precision * recall)/(precision + recall), where precision = TP/(TP + FP) and recall = TP/(TP + FN).
- Accuracy: Accuracy is the proportion of correct predictions made by the model out of the total number of predictions. It is calculated as (TP + TN)/(TP + TN + FP + FN).

In the calculations given for these performance metrics, TP, TN, FP, and FN are represented as True Positives, True Negatives, False Positives, and False Negatives respectively (Buckland and Gey, 1994; Tan and Zhu, 2023).

6.3. Results

In this study, a CNN and RNN-based deep network was developed for voice analysis. In the first stage of the experimental studies, transfer learning approaches (Fine-tuning and Training from scratch) based on the pre-trained AlexNet model were used. For this purpose, the first 16 layers of this deep model are frozen, and then fully connected, softmax, and classification layers are added. In the fine-tuning process, pre-trained weights are used, while in another approach, training is performed from scratch. In the training phase of these deep networks, the parameters given in Table 3 were used.

Confusion matrices obtained from Fine-tuning and Training from scratch approaches based on the pre-trained AlexNet model using the training parameters given in Table 5 are given in Fig. 8.

According to the data in Fig. 9, sensitivity, specificity, precision, F1-score, and accuracy values obtained with the Training from scratch approach were produced 86.14 %, 93.07 %, 86.96 %, 86.02 %, and 86.21 % values, respectively, while Fine-tuning approach was 91.32 %, 95.66 %, 91.70 %, 91.21 %, and 91.38 % values, respectively. Additionally, using the Fine-tuning approach, it recognized 100 % healthy people. As a result, it was observed that the fine-tuning process achieved 5 % higher performance than the training-from-scratch process. In this direction, using the learned weights of pre-trained deep architectures to solve another problem has been proven to provide high performance. Accordingly, it was decided to continue our experimental studies by using the fine-tuning process. In the second experimental study, five pre-trained frozen conv layers were given as separate inputs to the GRU and LSTM layers. Then, fully connected, ReLU and Softmax layers were added for each RNN model, and the training process was carried out. 94.83 % accuracy was obtained using the LSTM-based CNN model, while the GRU-based CNN model produced an accuracy score of 93.1 %.

Table 3
General characteristics of the proposed architecture.

Name	Filters	Type	Activations	Learnable Properties
Data	–	Image Input	227x227x3	–
Conv1	(96, 11x11) Stride [4 4] Padding [0 0 0 0]	2-D Convolution	55x55x96	Weights 11x11 × 3 × 96 Bias 1 × 1 × 96
Relu1	–	ReLU	55x55x96	–
Norm1	–	Cross Channel Normalization with 5 channels per element	55x55x96	–
Pool1	(3x3) Stride [2 2] Padding [0 0 0 0]	2-D Max Pooling	27x27x96	–
Conv2	2 groups of (128, 5x5) Stride [1 1] Padding [2 2 2 2]	2-D Grouped Convolution	27x27x256	Weights 5 × 5 × 48x128 Bias 1 × 1 × 128x2
Relu2	–	ReLU	27x27x256	–
Norm2	–	Cross Channel Normalization with 5 channels per element	27x27x256	–
Pool2	(3x3) Stride [2 2] Padding [0 0 0 0]	2-D Max Pooling	13x13x256	–
Conv3	(384, 3x3) Stride [1 1] Padding [1 1 1 1]	2-D Convolution	13x13x384	Weights 3 × 3 × 256x384 Bias 1 × 1 × 384
Relu3	–	ReLU	13x13x384	
Conv4	2 groups of (192, 3x3) Stride [1 1] Padding [1 1 1 1]	–	13x13x384	Weights 3 × 3 × 192x192x2 Bias 1 × 1 × 128x2
Relu4		ReLU	13x13x384	–
Conv5	2 groups of (128, 3x3) Stride [1 1] Padding [1 1 1 1]	–	13x13x256	Weights 3 × 3 × 192x128x2 Bias 1 × 1 × 128x2
Relu5	–	ReLU	13x13x256	–
Pool5	(3x3) Stride [2 2] Padding [0 0 0 0]	–	6 × 6 × 256	–
Flatten		Flatten	9216x1	–
gru	300 hidden units	GRU	300x1	InputWeights 900x9216 RecurrentWeights 900x300 Bias 900x1
Drop1	–	Dropout		–
Lstm	300 hidden units	LSTM	300x1	InputWeights 1200x9216 RecurrentWeights 1200x300 Bias 1200x1
Drop2	–	Dropout	300x1	–
Addition	–	Element-wise addition of 2 inputs	300x1	–
Fc1	1024	Fully Connected	1024x1	Weights 1024x300 Bias 1024x1
Relu6	–	ReLU	1024x1	
Fc2	256	Fully Connected	256x1	Weights 256x1024 Bias 256x1
Relu7	–	ReLU	256x1	
fc	3	Fully Connected	3x1	Weights 3x256 Bias 3x1
Softmax	–	Softmax	3x1	–
class output	–	Classification Output	3x1	–

Table 4
Individual distribution by group, age, and gender.

Group	Age Range	Gender	Number of Individuals	Total number of Individuals per category
MS	18–65	Male	47	99
MS	18–65	Female	52	
Other	18–65	Male	51	96
Other	18–65	Female	45	
Healthy	18–65	Male	39	95
Healthy	18–65	Female	56	
Total number of individuals				290

According to the results RNN-based approaches have achieved better results than the raw CNN model. Based on these results, confusion matrixes of both RNN-based CNN models are given in Fig. 9.

As can be seen in Fig. 9, LSTM and GRU-based models classified healthy and MS classes as 100 %, while other diseases were detected with an average performance of 81 %.

In the last phase of the experimental studies, performance results were calculated using the proposed CNN-based architecture in which GRU and LSTM models are connected in parallel. The confusion matrix obtained from this experimental study is given in Fig. 10.

As can be seen from Fig. 10, sensitivity, specificity, precision, F1-score, and accuracy values obtained with the proposed model were produced 96.49 %, 98.25 %, 96.97 %, 96.56 %, and 96.55 % values, respectively. According to these results, the proposed model achieved superior success compared to the results obtained from all other experimental studies. In addition, the proposed model classified the healthy and MS classes as 100 %, while it incorrectly classified only 2 test data in other diseases. The results of the methods, including AlexNet from scratch, fine-tuned AlexNet, LSTM based CNN, GRU based CNN, and the proposed method, using the same dataset, are compared in Table 6. Overall, the Proposed Method achieved the highest accuracy, precision, recall, and F1 score for all three classes. The LSTM based CNN and GRU based CNN methods also performed well, but the Proposed Method was slightly better. The AlexNet methods performed the worst, especially for the Other class. The results show that the Proposed Method is a promising method for this classification task. It is able to accurately classify all three classes, even for the difficult Other class.

In evaluating the study's limitations, the parallel integration of LSTM and GRU introduces increased computational complexity and memory demands, which may constrain the model's scalability. The model's performance is significantly contingent on the quality and quantity of the data, potentially affecting its generalizability to other datasets. Furthermore, there is a persistent risk of overfitting, particularly if appropriate regularization techniques are not implemented. Fig. 11

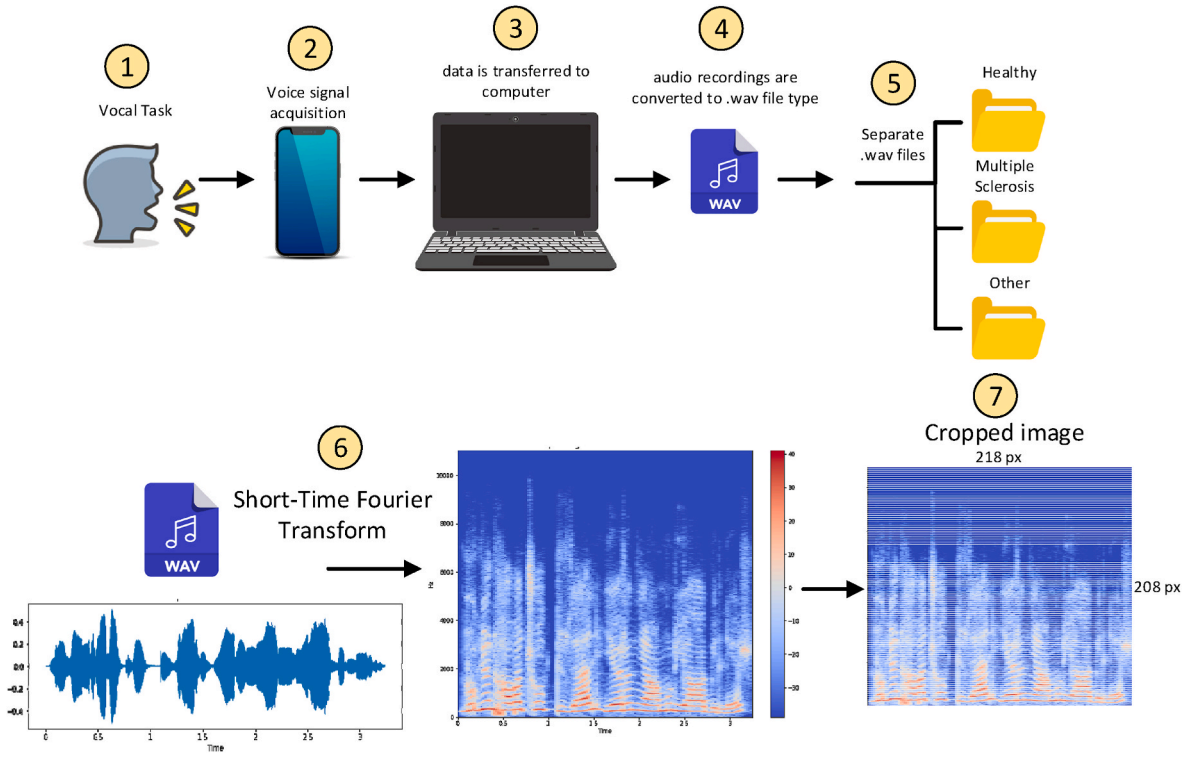

Fig. 7. Dataset preparation process.

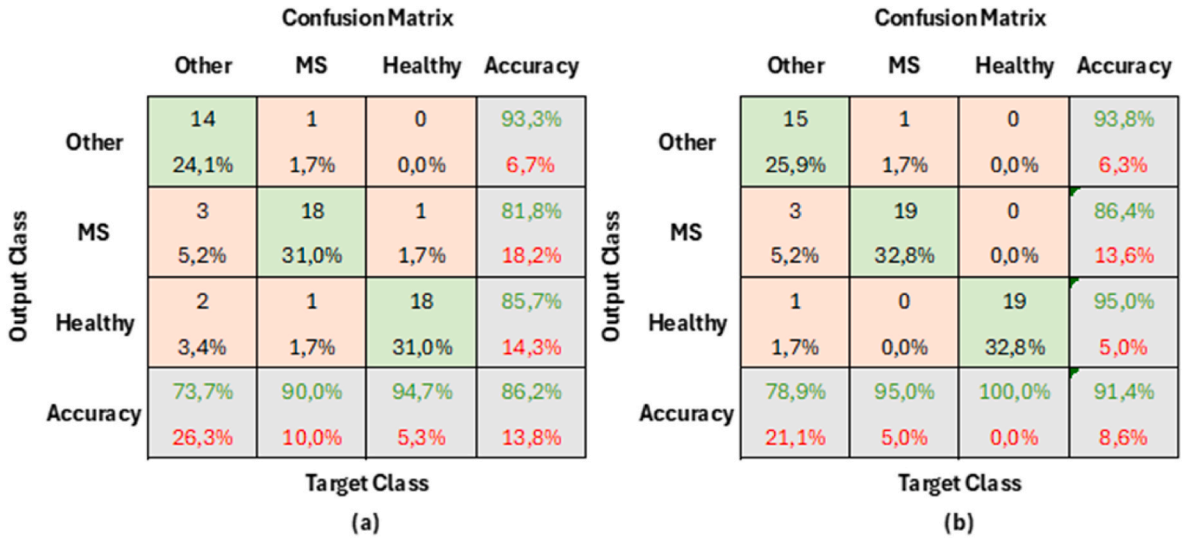

Fig. 8. Confusion matrix's, a) Training from scratch, b) Fine-tuning.

Table 5
The deep parameters used in the training phase.

Parameters	Values
Epoch	50
Batch Size	16
InitialLearnRate	1e-4
LearnRateDropPeriod	2
L2Regularization	5e-4
Optimization method	Adam

presents the overall accuracy of different models, highlighting the performance of the proposed method compared to other approaches in sound classification.

On the other hand, to better evaluate the generalizability of the proposed architecture, a 5-fold cross-validation method is used. This method allows for a more reliable performance evaluation over the entire dataset by training the model on each subset. Cross-validation provides an opportunity to observe how the model performs on different data subsets, which minimizes the risk of model overfitting. In addition, the use of cross-validation reduces the model's dependence on the training data, which increases its generalizability to real-world conditions. Table 7 shows the accuracy scores of the fine-tuned AlexNet, the LSTM-based CNN, the GRU-based CNN, and the proposed model

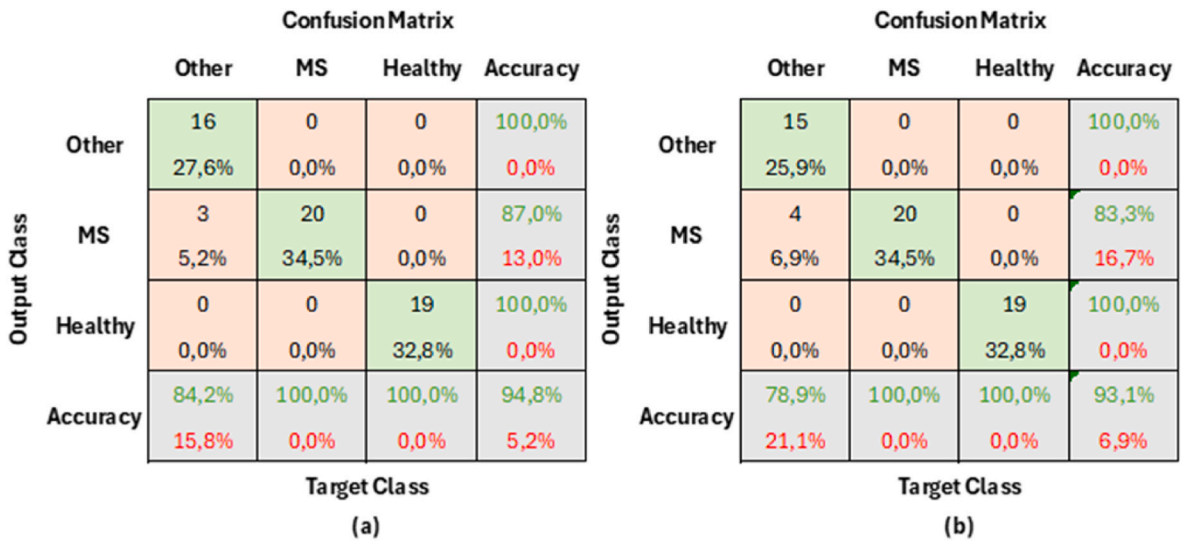

Fig. 9. Confusion matrixes of RNN-based CNN models, a) LSTM, b) GRU.

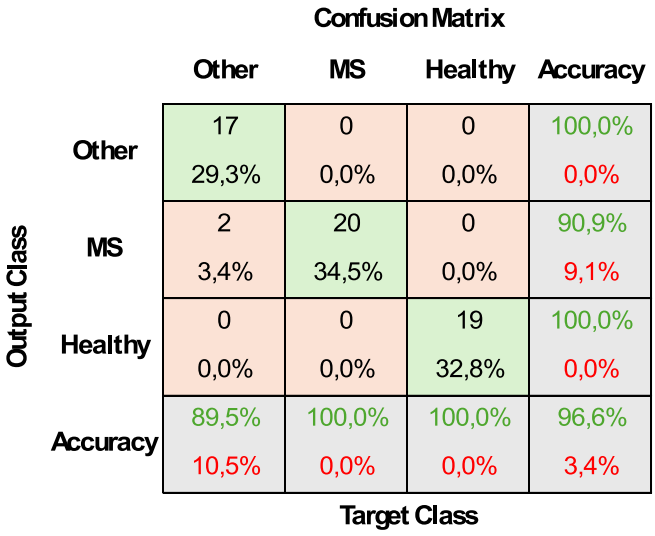

Fig. 10. Confusion matrix of proposed model.

at each layer. When the 5-fold cross-validation results are analyzed, the overall accuracy of the proposed model is significantly superior to the other models. These results show that the model provides high accuracy, precision, and generalizability, and performs consistently on different datasets.

The results in Table 7 show that the proposed model has the highest performance in terms of overall accuracy across all layers, which supports the consistent success of the model on the dataset. These results show that cross-validation is an important method for evaluating model accuracy, and the proposed model has greater generalizability than other approaches. Furthermore, in the previous experiments, 80 % of the spectral images were used as training data and 20 % of the spectral images were used as test data. In these settings, the proposed model is also found to perform well, which reinforces the accuracy and generalizability of the model.

7. Conclusions

In this study, an AI-based method was developed to assist in the diagnosis of neurological diseases through voice recordings. The main contributions include the use of CNN and RNN-based deep networks, which proved effective and reliable for accurately detecting nervous system diseases by analyzing audio signals. The proposed hybrid model, combining CNN and RNN architectures, enabled more effective analysis and classification of these signals. The dataset was expanded to cover a wide range of neurological conditions, enhancing the model's ability to describe various diseases and broadening its clinical applicability. The study highlighted the potential of voice analysis and AI-based approaches, emphasizing their importance in early detection and clinical

Table 6
Comparison of the methods used.

Method	Class	n truth	n classified	Accuracy	Precision	Recall	F1 Score	Overall Accuracy
AlexNet from Stratch	Other	19	15	89.66 %	0.93	0.74	0.82	86.21 %
	MS	20	22	89.66 %	0.82	0.90	0.86	
	Healthy	19	21	93.1 %	0.86	0.95	0.90	
Fine tuned AlexNet	Other	19	16	91.38 %	0.94	0.79	0.86	91.38 %
	MS	20	22	93.1 %	0.86	0.95	0.90	
	Healthy	19	20	98.28 %	0.95	1	0.97	
LSTM based CNN	Other	19	16	94.83 %	1.0	0.84	0.91	94.83 %
	MS	20	23	94.83 %	0.87	1.0	0.83	
	Healthy	19	19	100 %	1.0	1.0	1.0	
GRU based CNN	Other	19	15	93.1	1.0	0.79	0.88	93.1 %
	MS	20	24	93.1	0.83	1.0	0.91	
	Healthy	19	19	100 %	1.0	1.0	1.0	
Proposed Method	Other	19	17	96.55 %	1.0	0.89	0.94	96.55 %
	MS	20	22	96.55 %	0.91	1.0	0.95	
	Healthy	19	19	100 %	1.0	1.0	1.0	

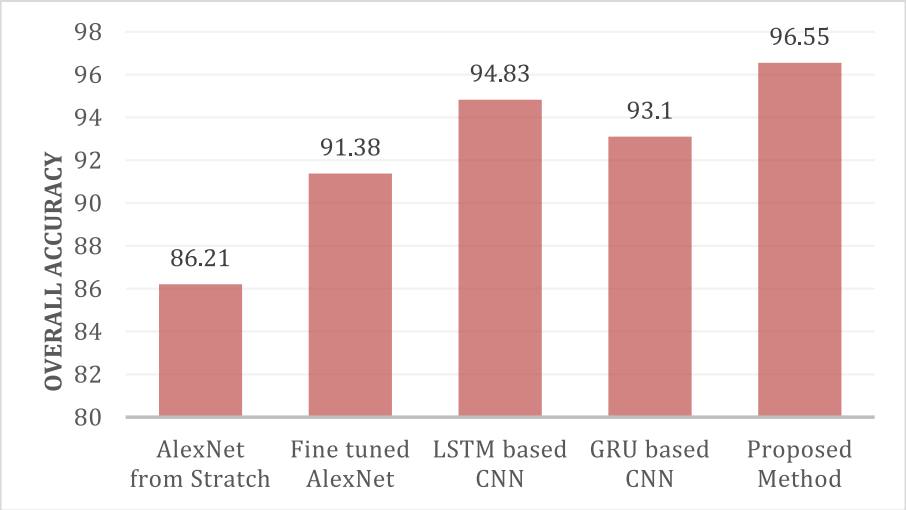

Fig. 11. Comparison of model performance in sound classification.

Table 7
Performance evaluation (%) of models using 5-fold cross-validation.

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	All
Fine-tuned AlexNet	91.23	91.38	87.93	94.74	87.72	90.60
LSTM based CNN	91.23	93.10	93.10	94.74	92.98	93.03
GRU Based CNN	87.72	93.10	91.38	96.49	91.23	91.98
Proposed Model	91.23	96.55	94.83	96.49	94.74	94.77

applications, particularly as AI advancements now allow computers to detect subtle vocal changes that may be indicative of neurological issues, which might go unnoticed by the human ear.

The model's high accuracy has the potential to greatly enhance clinical workflows by improving diagnostic accuracy, optimizing decision support, and enabling better patient monitoring. However, to ensure its successful adoption in healthcare environments, future validation steps must focus on diverse data, clinical testing, integration with existing systems, and addressing regulatory and ethical concerns. These steps will help transition the model from a research tool to a trusted and effective component of clinical practice.

The results of the developed deep network system have the potential to provide diagnostic support in neurological applications. The use of voice recordings represents a non-invasive approach that offers the possibility of diagnosing patients without causing discomfort. This study provides a significant foundation for future research in AI-based neurological disease diagnosis, emphasizing the need to expand the model's scope and enhance its general applicability. Future research can focus on improving the performance and efficiency of CNN and RNN algorithms, exploring practical applications in clinical settings, and ensuring the protection of personal health data. These efforts are crucial for advancing early diagnosis and treatment of neurological diseases and making a meaningful contribution to healthcare.

Our study highlights the potential of voice analysis in the classification of neurological disorders, and an important direction for future research is testing the model on larger datasets and in multi-center settings to ensure its robustness and clinical applicability. However, one limitation is that the study does not explicitly address how the model performs across diverse populations, such as non-Turkish speakers, or in real-world clinical environments. To enhance generalizability, future research should focus on validating the model in linguistically and geographically diverse cohorts. This would help assess its adaptability to different languages, accents, and clinical conditions, ultimately improving its applicability in global healthcare settings.

CRediT authorship contribution statement

Emel Soylu: Writing – review & editing, Writing – original draft, Visualization, Supervision, Software, Project administration, Methodology, Investigation, Conceptualization. **Sema Gül:** Supervision, Data curation, Conceptualization. **Kübra Aslan Koca:** Writing – review & editing, Writing – original draft, Visualization, Software, Data curation, Conceptualization. **Muammer Türkoğlu:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Methodology, Conceptualization. **Murat Terzi:** Writing – original draft, Supervision, Formal analysis, Data curation, Conceptualization. **Abdulkadir Şengür:** Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Conceptualization.

Declaration of competing interest

I, Emel SOYLU hereby declare that I have no conflicts of interest to disclose regarding the publication "Speech Signal-based Accurate Neurological Disorders Detection Using CNN and RNN-Based Deep Network"

We have not received any financial support or incentives from any organization that could influence the content of this publication. Additionally, I have no personal or professional relationships that could bias my interpretation or presentation of the research findings.

We affirm that all information presented in this publication is accurate to the best of our knowledge and has been obtained through ethical research practices. Any opinions expressed in this publication are solely our own and do not reflect the views of any affiliated institutions or organizations.

Should any potential conflicts of interest arise in the future that may affect the integrity of this publication, I am committed to promptly disclosing them to the relevant parties.

Abbreviations

ADC	Analog-To-Digital Converter
AI	Artificial Intelligence
ANN	Artificial neural network
CNN	Convolutional Neural Networks
GRU	Gated Recurrent Unit
LSTM	Long Short-Term Memory
MFCC	Mel-Frequency Cepstral Coefficient
MS	Multiple Sclerosis
RMT	Respiratory Muscle Training

RNN	Recurrent Neural Networks
STFT	Short-Time Fourier Transform References
ReLU	Rectified Linear Unit

Data availability

The data substantiating the conclusions of this research can be obtained from the corresponding author upon making a reasonable request.

References

Adamec, I., Barun, B., Gabelić, T., et al., 2013. Delay in the diagnosis of multiple sclerosis in Croatia. Clin. Neurol. Neurosurg. 115, S70–S72.

Akkem, Y., Biswas, S.K., Varanasi, A., 2023a. Smart farming using artificial intelligence: a review. Eng. Appl. Artif. Intell. 120, 105899. https://doi.org/10.1016/j.engappai.2023.105899.

Akkem, Y., Kumar, B.S., Varanasi, A., 2023b. Streamlit application for advanced Ensemble learning methods in crop recommendation systems – a review and implementation. Indian J. Sci. Technol. 16, 4688–4702. https://doi.org/10.17485/ijst/v16i48.2850.

Akkem, Y., Biswas, S.K., Varanasi, A., 2024. A comprehensive review of synthetic data generation in smart farming by using variational autoencoder and generative adversarial network. Eng. Appl. Artif. Intell. 131, 107881. https://doi.org/10.1016/j.engappai.2024.107881.

Alghamdi, N.S., Zakariah, M., Hoang, V.T., Elahi, M.M., 2022. Neurogenerative disease diagnosis in cepstral domain using MFCC with deep learning. Comput. Math. Methods Med. 2022.

Almeida, J.S., Rebouças Filho, P.P., Carneiro, T., et al., 2019. Detecting Parkinson's disease with sustained phonation and speech signals using machine learning techniques. Pattern Recognit. Lett. 125, 55–62.

Arnold, R., Gaskill, C.S., Bausek, N., 2023. Effect of Combined Respiratory Muscle Training (cRMT) on Dysphonia Following Single CVA: A Retrospective Pilot Study. J Voice. https://doi.org/10.1016/j.jvoice.2021.03.014.

Asfiati, S., Riky, M.N., Rajagukguk, J., others, 2020. Measurement and evaluation of sound intensity at the Medan Railway Station using a sound level meter. J. Phys. Conf., 12063

Asiaee, M., Vahedian-Azimi, A., Atashi, S.S., et al., 2022. Voice quality evaluation in patients with COVID-19: an acoustic analysis. J. Voice 36, 879–e13.

Ball, L.J., Beukelman, D.R., Pattee, G.L., 2002. Timing of speech deterioration in people with amyotrophic lateral sclerosis. J. Med. Speech Lang. Pathol. 10, 231–236.

Becker, G., Müller, A., Braune, S., et al., 2002. Early diagnosis of Parkinson's disease. J. Neurol. 249, iii40–iii48.

Bezoui, M., Elmoutaouakkil, A., Beni-hssane, A., 2016. Feature extraction of some Quranic recitation using mel-frequency cepstral coeficients (MFCC). In: 2016 5th International Conference on Multimedia Computing and Systems (ICMCS), pp. 127–131.

Bilgera, C., Yamamoto, A., Sawano, M., et al., 2018. Application of convolutional long short-term memory neural networks to signals collected from a sensor network for Autonomous gas Source localization in outdoor environments. Sensors. https://doi.org/10.3390/s18124484.

Bouafif, L., Ellouze, N., 2018. A new E-health tool for early identification of voice and neurological pathologies by speech processing. Int. J. Adv. Comput. Sci. Appl. https://doi.org/10.14569/ijacsa.2018.090865.

Brabenec, L., Mekyska, J., Galaz, Z., Rektorova, I., 2017. Speech disorders in Parkinson's disease: early diagnostics and effects of medication and brain stimulation. J. Neural Transm. 124, 303–334.

Brooks, D.J., 1998. The early diagnosis of Parkinson's disease. Ann. Neurol. 44, S10–S18.

Buckland, M., Gey, F., 1994. The relationship between recall and precision. J. Am. Soc. Inf. Sci. 45, 12–19.

Cárdenas-Robledo, S., Lopez-Reyes, L., Arenas-Vargas, L.E., et al., 2021. Delayed diagnosis of multiple sclerosis in a low prevalence country. Neurol. Res. 43, 521–527.

Chen, C., Moro-Velazquez, L., Ozbolt, A.S., et al., 2022. Phonatory analysis on Parkinson's disease patients attending singing and discussion therapy (parkinsonics) using signal processing techniques. In: 2022 IEEE Signal Processing in Medicine and Biology Symposium (SPMB), pp. 1–5.

Chen, Z., Zhu, P., Qiu, W., Guo, J., Li, Y., 2023. Deep learning in automatic detection of dysphonia: Comparing acoustic features and developing a generalizable framework. Int. J. Lang. Commun. Disord. 58 (2), 279–294.

Cho, K., Van Merriënboer, B., Gulcehre, C., et al., 2014. Learning Phrase Representations Using RNN Encoder-Decoder for Statistical Machine Translation arXiv Prepr arXiv14061078.

Chu, L.W., others, 2012. Alzheimer's disease: early diagnosis and treatment. Hong Kong Med J 18, 228–237.

Cnockaert, L., Schoentgen, J., Auzou, P., et al., 2008. Low-Frequency Vocal Modulations in Vowels Produced by Parkinsonian Subjects. Speech Commun. https://doi.org/10.1016/j.specom.2007.10.003.

Deng, M., Meng, T., Cao, J., et al., 2020. Heart sound classification based on improved MFCC features and convolutional recurrent neural networks. Neural Networks 130, 22–32.

Duffy, J.R., others, 2012. Motor Speech Disorders: Substrates, Differential Diagnosis, and Management. Elsevier Health Sciences.

Faust, O., Hagiwara, Y., Hong, T.J., et al., 2018. Deep learning for healthcare applications based on physiological signals: a review. Comput Methods Programs Biomed. https://doi.org/10.1016/j.cmpb.2018.04.005.

Feijó, A.V., Parente, M.A., Behlau, M., et al., 2004. Acoustic analysis of voice in multiple sclerosis patients. J Voice 18, 341–347.

Freed, D.B., 2023. Motor Speech Disorders: Diagnosis and Treatment. plural publishing.

Frid, A., Safra, E.J., Hazan, H., et al., 2014. Computational diagnosis of Parkinson's disease directly from natural speech using machine learning techniques. In: 2014 IEEE International Conference on Software Science, Technology and Engineering, pp. 50–53.

Gilchrist, N.H., 1980. Analogue-to-Digital and digital-to-analogue converters for high quality sound. In: Audio Engineering Society Convention, vol. 65.

Goberman, A.M., Coelho, C., 2002. Acoustic analysis of Parkinsonian speech I: speech characteristics and L-Dopa therapy. NeuroRehabilitation 17, 237–246.

Gökçen, A., 2021. Computer-aided diagnosis system for chronic obstructive pulmonary disease using empirical wavelet transform on auscultation sounds. Comput J 64, 1775–1783.

Gosztolya, G., Tóth, L., Svindt, V., et al., 2022. Using acoustic deep neural network embeddings to detect multiple sclerosis from speech. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6927–6931.

Group, C.M.D., Campbell, P., Rooney, S., et al., 1996. Speech and language therapy interventions for speech problems in Parkinson's disease. Cochrane Database Syst Rev 2022.

Gullapalli, A.S., Mittal, V.K., 2021. Early Detection of Parkinson's Disease through Speech Features and Machine Learning: A Review. https://doi.org/10.1007/978-981-16-4177-0_22.

Hadders-Algra, M., 2014. Early diagnosis and early intervention in cerebral palsy. Front Neurol 5, 185.

Haq, A.U., Li, J., Memon, M.H., et al., 2018. Comparative analysis of the classification performance of machine learning classifiers and deep neural network classifier for prediction of Parkinson disease. In: 2018 15th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP), pp. 101–106.

Hassan, M.M., Tahir, M.H., Ameeq, M., et al., 2023. Risk factors identification of COVID-19 patients with chronic obstructive pulmonary disease: a retrospective study in Punjab-Pakistan. Immunity. Inflamm Dis 11, e981.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput 9, 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.

Hong, X., Lin, R., Yang, C., et al., 2019. Predicting Alzheimer's disease using LSTM. Ieee Access. https://doi.org/10.1109/access.2019.2919385.

Huzaifah, M., 2017. Comparison of Time-Frequency Representations for Environmental Sound Classification Using Convolutional Neural Networks arXiv Prepr arXiv170607156.

Igra, M.S., Paling, D., Wattjes, M.P., et al., 2017. Multiple sclerosis update: use of MRI for early diagnosis, disease monitoring and assessment of treatment related complications. Br J Radiol 90, 20160721.

Islam, M.A., Bandyopadhyaya, I., Bhattacharyya, P., Saha, G., 2018. Multichannel lung sound analysis for asthma detection. Comput Methods Programs Biomed 159, 111–123.

Jin, B., Xu, X., 2024a. Pre-owned Housing Price Index Forecasts Using Gaussian Process Regressions.

Jin, B., Xu, X., 2024b. Price forecasting through neural networks for crude oil, heating oil, and natural gas. Meas Energy 1, 100001. https://doi.org/10.1016/j.meaene.2024.100001.

Jin, B., Xu, X., 2024c. Wholesale Price Forecasts of Green Grams Using the Neural Network. Asian J Econ Bank. https://doi.org/10.1108/ajeb-01-2024-0007.

Jin, B., Xu, X., 2024d. Forecasting wholesale prices of yellow corn through the Gaussian process regression. Neural Comput Appl 36, 8693–8710. https://doi.org/10.1007/s00521-024-09531-2.

Jin, B., Xu, X., 2024e. Machine learning predictions of regional steel price indices for east China. Ironmak \& Steelmak, 03019233241254891.

Jin, B., Xu, X., 2024f. Palladium price predictions via machine learning. Mater Circ Econ 6. https://doi.org/10.1007/s42824-024-00123-y.

Jin, B., Xu, X., 2024g. Gaussian Process Regression Based Silver Price Forecasts. J Uncertain Syst.

K, U.R., Holi, M.S., 2015. A hybrid model for neurological disordered voice classification using time and frequency domain features. Artif Intell Res. https://doi.org/10.5430/air.v5n1p87.

Kapur, A., Sarawgi, U., Wadkins, E., et al., 2020. Non-invasive silent speech recognition in multiple sclerosis with dysphonia. In: Machine Learning for Health Workshop, pp. 25–38.

Karlsson, F., Hartelius, L., 2021. On the primary influences of age on articulation and phonation in maximum performance tasks. Languages. https://doi.org/10.3390/languages6040174.

Kennedy, P., 2013. Impact of delayed diagnosis and treatment in clinically isolated syndrome and multiple sclerosis. J Neurosci Nurs 45, S3–S13.

Kent, R.D., Rosenbek, J.C., 1982. Prosodic disturbance and neurologic lesion. Brain Lang 15, 259–291.

Khamparia, A., Gupta, D., Nguyen, N.G., et al., 2019. Sound classification using convolutional neural network and tensor deep stacking network. IEEE Access 7, 7717–7727.

Kim, M.J., Cao, B., An, K., Wang, J., 2018. Dysarthric speech recognition using convolutional LSTM neural network. https://doi.org/10.21437/interspeech.2018-2250.
Laakso, M.P., Hallikainen, M., Hänninen, T., et al., 2000. Diagnosis of Alzheimer's disease: MRI of the hippocampus vs delayed recall. Neuropsychologia 38, 579–584.
Lee, K.-H., He, X., Zhang, L., Yang, L., 2018. Cleannet: transfer learning for scalable image classifier training with label noise. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5447–5456.
Lehman, V.T., Black, D.F., DeLone, D.R., et al., 2020. Current concepts of cross-sectional and functional anatomy of the cerebellum: a pictorial review and atlas. Br J Radiol 93, 20190467.
Li, B., 2011. On identity authentication technology of distance education system based on voiceprint recognition. In: Proceedings of the 30th Chinese Control Conference, pp. 5718–5721.
Li, S., Li, F., Tang, S., Xiong, W., 2020. A review of computer-aided heart sound detection techniques. Biomed Res Int 2020.
Liang, G., Zheng, L., 2020. A transfer learning method with deep residual network for pediatric pneumonia diagnosis. Comput Methods Programs Biomed 187, 104964.
Liu, L., Zhao, S., Chen, H., Wang, A., 2020. A new machine learning method for identifying Alzheimer's disease. Simul Model Pract Theory 99, 102023.
Lopes, L.W., Silva, HF da, Evangelista, D. da S., et al., 2016. Relationship between vocal symptoms, severity of voice disorders, and laryngeal diagnosis in patients with voice disorders. CoDAS 439–445.
Mahmood, A., Khan, M.M., Imran, M., et al., 2023. End-to-End deep learning method for detection of invasive Parkinson's disease. Diagnostics. https://doi.org/10.3390/diagnostics13061088.
Mahmoud, A.S., Lamouchi, O., Belghith, S., 2024. Advancements in machine learning and deep learning for early diagnosis of chronic kidney diseases: a comprehensive review. Babylonian J Mach Learn 2024, 149–156.
Maiti, S., 2024. Deep learning based prediction and monitoring of Parkinson's disease using voice data. https://doi.org/10.21203/rs.3.rs-4698818/v1.
Majda-Zdancewicz, E., Potulska-Chromik, A., Nojszewska, M., Kostera-Pruszczyk, A., 2024. Speech signal analysis in patients with Parkinson's disease, taking into Account phonation, articulation, and prosody of speech. Appl Sci 14, 11085.
Mauch, M., Cannam, C., Bittner, R., et al., 2015. Computer-aided Melody Note Transcription Using the Tony Software: Accuracy and Efficiency.
Miller, J.R., 2004. The importance of early diagnosis of multiple sclerosis. J Manag care Pharm JMCP 10, S4–11.
Moumdjian, L., Six, J., Veldkamp, R., et al., 2022. Embodied learning in multiple sclerosis using melodic, sound, and visual feedback: a potential rehabilitation approach. Ann N Y Acad Sci 1513, 153–169.
Muneeb Hassan, M., Ameeq, M., Jamal, F., et al., 2023. Prevalence of covid-19 among patients with chronic obstructive pulmonary disease and tuberculosis. Ann Med 55, 285–291.
Nabih-Ali, M., El-Dahshan, E.-S.A., Yahia, A.S., 2017. A review of intelligent systems for heart sound signal analysis. J Med Eng \& Technol 41, 553–563.
Nagawade, M.S., Ratnaparkhe, V.R., 2017. Musical instrument identification using MFCC. In: 2017 2nd IEEE International Conference on Recent Trends in Electronics, Information \& Communication Technology (RTEICT), pp. 2198–2202.
Ngo, Q.C., Motin, M.A., Pah, N.D., et al., 2022. Computerized analysis of speech and voice for Parkinson's disease: a systematic review. Comput Methods Programs Biomed, 107133.
Nguyen-Da, T., Li, Y., Peng, C.-L., et al., 2023. Tourism demand prediction after COVID-19 with deep learning hybrid CNN–LSTM—case study of vietnam and provinces. Sustainability. https://doi.org/10.3390/su15097179.
Noffs, G., Boonstra, F.M.C., Perera, T., et al., 2020. Acoustic speech analytics are predictive of cerebellar dysfunction in multiple sclerosis. The Cerebellum 19, 691–700.
Norel, R., Pietrowicz, M., Agurto, C., et al., 2018. Detection of amyotrophic lateral sclerosis (ALS) via acoustic analysis. bioRxiv, 383414.
Novotny, M., Melechovsky, J., Rozenstoks, K., et al., 2020. Comparison of automated acoustic methods for oral diadochokinesis assessment in amyotrophic lateral sclerosis. J Speech, Lang Hear Res 63, 3453–3460.
Nzwalo, H., de Abreu, D., Swash, M., et al., 2014. Delayed diagnosis in ALS: the problem continues. J Neurol Sci 343, 173–175.
Orozco-Arroyave, J.R., Belalcazar-Bolanos, E.A., Arias-Londoño, J.D., et al., 2015. Characterization methods for the detection of multiple voice disorders: neurological, functional, and laryngeal diseases. IEEE J Biomed Heal informatics 19, 1820–1828.
O'Shea, K., Nash, R., 2015. An Introduction to Convolutional Neural Networks arXiv Prepr arXiv151108458.
Pah, N.D., Motin, M.A., Kumar, D.K., 2022. Voice analysis for diagnosis and monitoring Parkinson's disease. Tech Assess Park Diagnosis Rehabil 119–133.
Pasnau, R., 1999. What is sound? Philos Q 49, 309–324.
Philip, A.G.S., Hewitt, J.R., 1980. Early diagnosis of neonatal sepsis. Pediatrics 65, 1036–1041.
Piczak, K.J., 2015. Environmental sound classification with convolutional neural networks. In: 2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP), pp. 1–6.
Prabakaran, D., Sriuppili, S., 2021. Speech processing: MFCC based feature extraction techniques-an investigation. Journal of Physics: Conference Series, 12009.
Prasad, B.R., Deepa, N., 2021. Classification of analyzed text in speech recognition using RNN-LSTM in comparison with convolutional neural network to improve precision for identification of keywords. Rev Gestão Inovação E Tecnol. https://doi.org/10.47059/revistageintec.v11i2.1739.
Purwins, H., Li, B., Virtanen, T., et al., 2019. Deep learning for audio signal processing. IEEE J Sel Top Signal Process 13, 206–219.
Rahman, S., Hasan, M.M., Sarkar, A.K., Khan, F., 2023. Classification of Parkinson's disease using speech signal with machine learning and deep learning approaches. Eur J Electr Eng Comput Sci. https://doi.org/10.24018/ejece.2023.7.2.488.
Ren, Z., Nguyen, T.T., Nejdl, W., 2022. Prototype learning for interpretable respiratory sound analysis. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 9087–9091.
Renauld, S., Mohamed-Sa, id, L., Macoir, J., 2016. Language disorders in multiple sclerosis: a systematic review. Mult Scler Relat Disord 10, 103–111.
Richards, D., Morren, J.A., Pioro, E.P., 2020. Time to diagnosis and factors affecting diagnostic delay in amyotrophic lateral sclerosis. J Neurol Sci 417, 117054.
Rosen, K.M., Kent, R.D., Duffy, J.R., 2005. Task-based profile of vocal intensity decline in Parkinson's disease. Folia Phoniatr Logop. https://doi.org/10.1159/000081959.
Rusz, J., Cmejla, R., Ruzickova, H., Ruzicka, E., 2011. Quantitative acoustic measurements for characterization of speech and voice disorders in early untreated Parkinson's disease. J Acoust Soc Am 129, 350–367.
Şakar HG Evrişimli Sinir Ağları (Convolutional Neural Networks—CNN). In: Medium. https://hilalgozutok.medium.com/evrişimli-sinir-ağları-convolutional-neural-networks-cnn-e61470e9bdb1.
Santosh, K.C., Rasmussen, N., Mamun, M., Aryal, S., 2022. A systematic review on cough sound analysis for Covid-19 diagnosis and screening: is my cough sound COVID-19? PeerJ Comput Sci 8, e958.
Sharma, G., Umapathy, K., Krishnan, S., 2020. Trends in audio signal feature extraction methods. Appl Acoust 158, 107020.
Sheela, M.S., Amirthayogam, G., Hephzipah, J.J., et al., 2024. Advanced brain tumor classification using DEEPBELEIF-CNN method. Babylonian J Mach Learn 2024, 89–101.
Shi, L., Ahmad, I., He, Y., Chang, K., 2018. Hidden Markov model based drone sound recognition using MFCC technique in practical noisy environments. J Commun Networks 20, 509–518.
Stft. https://musicinformationretrieval.com/stft.html.
Su, Y., Zhang, K., Wang, J., Madani, K., 2019. Environment sound classification using a two-stream CNN based on decision-level fusion. Sensors 19, 1733.
Süzen, A.A., Yildiz, Z., Yilmaz, T., 2019. LSTM Tabanlı Derin Sinir Ağı Ile Ayak Taban Basınç Verilerinden VKİ Durumlarının Sınıflandırılması. Bitlis Eren Üniversitesi Fen Bilim Derg. https://doi.org/10.17798/bitlisfen.540273.
Syed, Z.S., Memon, S.A., Memon, A.L., 2020. Deep Acoustic Embeddings for Identifying Parkinsonian Speech. Int J Adv Comput Sci Appl. https://doi.org/10.14569/ijacsa.2020.0111089.
Tan, S.C., Zhu, S., 2023. Binary search of the optimal cut-point value in ROC analysis using the F1 score. J Physics: Conference Series, 12002.
Tokozume, Y., Ushiku, Y., Harada, T., 2017. Learning from Between-Class Examples for Deep Sound Recognition arXiv Prepr arXiv171110282.
van Prooije, T., Knuijt, S., Oostveen, J., et al., 2024. Perceptual and acoustic analysis of speech in spinocerebellar ataxia type 1. Cerebellum 23, 112–120.
Verde, L., Pietro, G De, Sannino, G., 2018. Voice disorder identification by using machine learning techniques. Ieee Access. https://doi.org/10.1109/access.2018.2816338.
Wang, T.V., Song, P.C., 2022. Neurological voice disorders: a review. Int J Head Neck Surg 13, 32–40.
Wang, Q., Liu, Q., Xia, R., et al., 2020. Defect depth determination in laser infrared thermography based on LSTM-RNN. Ieee Access. https://doi.org/10.1109/access.2020.3018116.
Xiao, B., Xu, Y., Bi, X., et al., 2019. Follow the sound of children's heart: a deep-learning-based computer-aided pediatric CHDs diagnosis system. IEEE Internet Things J 7, 1994–2004.
Xu, X., Zhang, Y., 2021. Corn cash price forecasting with neural networks. Comput Electron Agric 184, 106120. https://doi.org/10.1016/j.compag.2021.106120.
Yalachkov, Y., Bergmann, H.J., Soyda\cs, D., et al., 2019. Cognitive impairment in multiple sclerosis is reflected by increased susceptibility to the sound-induced flash illusion. Front Neurol 10, 373.
Yamamoto, D., Arimura, H., Kakeda, S., et al., 2010. 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 34, 404–413.
Ye, F., Yang, J., 2021. A deep neural network model for speaker identification. Appl Sci 11, 3603.
Yu, Y., Si, X., Hu, C., Zhang, J., 2019. A review of recurrent neural networks: LSTM cells and network architectures. Neural Comput 31, 1235–1270.
Zahid, L., Maqsood, M., Durrani, M.Y., et al., 2020. A spectrogram-based deep feature assisted computer-aided diagnostic system for Parkinson's disease. IEEE Access 8, 35482–35495.
Zhang, X., Zou, Y., Shi, W., 2017a. Dilated convolution neural network with LeakyReLU for environmental sound classification. In: 2017 22nd International Conference on Digital Signal Processing (DSP), pp. 1–5.
Zhang, R., Tao, H., Wu, L., Guan, Y., 2017b. Transfer learning with neural networks for bearing fault diagnosis in changing working conditions. IEEE Access 5, 14347–14357.
Zhuang, F., Qi, Z., Duan, K., et al., 2020. A comprehensive survey on transfer learning. Proc IEEE 109, 43–76.