DEEP LEARNING TECHNIQUES FOR IMAGE AND SPEECH RECOGNITION: CURRENT TRENDS AND FUTURE DIRECTIONS

Soleman
Universitas Borobudur, Indonesia
*) corresponding author: solemediagrafik@gmail.com

Keywords

Deep Learning, Image Recognition, Transformer Model

Abstract

This study examines the latest developments and future directions of deep learning techniques in image and sound recognition. The study focuses on the analysis of various neural network architectures such as Convolutional Neural Networks (CNNs) for image processing and Recurrent Neural Networks (RNNs) for speech recognition. The methodology used includes a comprehensive literature study of the latest studies, evaluation of the performance of various models, and comparative analysis of existing techniques. The results showed a significant improvement in recognition accuracy, with CNNs achieving up to 98% accuracy for image classification and transformer-based models outperforming traditional RNNs in speech recognition. The challenges identified include high computational requirements, reliance on quality datasets, and model interpretability issues. The study also proposes several future development directions, including the integration of attention mechanisms, hybrid architectures, and more efficient learning techniques. In conclusion, despite the rapid progress, there is still significant room for innovation in improving the efficiency and reliability of deep learning-based image and voice recognition systems.

1. INTRODUCTION

The development of deep learning technology has brought about a revolution in data processing and automated decision-making, especially in the field of image and voice recognition (Mehrish et al., 2023). Deep learning, which is part of machine learning, uses artificial neural network architectures to study complex patterns in big data. With this capability, deep learning enables object and sound recognition with an increasingly high level of accuracy (Al-Fraihat et al., 2024). For example, convolutional neural networks (CNNs) have accelerated advances in image recognition, allowing computers to identify and classify objects in images or videos with great accuracy (Z. Zhang et al., 2018). Meanwhile, in the field of speech recognition, models such as repetitive neural networks (RNNs) and transformers have enabled computers to understand and translate human speech in real time. Both technologies have paved the way for a wide range of applications, from facial recognition in security systems to intelligent voice assistants in everyday devices, such as mobile phones and smart home systems (Khalil et al., 2019).

The contribution of deep learning technology to the advancement of information technology is significant because it expands the limits of machine capabilities to interact with the real world (Dargan et al., 2020). This technology has driven the development of AI-based products and services that not only help individuals in their daily activities but also support industry sectors such as healthcare, transportation, and education. In the health sector (Delić et al., 2019), for example, deep learning is used in medical image analysis to detect diseases earlier and with higher accuracy, thereby improving

diagnosis and treatment outcomes. In the transportation sector, this technology plays a role in object recognition systems in autonomous vehicles that are able to recognize road conditions and respond to situations quickly and accurately. As such, deep learning not only contributes to operational efficiency, but also to the safety and convenience of technology users (Delić et al., 2019).

Image and voice recognition plays a crucial role in supporting technological advancements and improving interaction between humans and machines. In the field of security, image recognition is widely used in surveillance and access control systems, such as in facial recognition technology in public areas and offices. With the ability to accurately recognize and identify faces, this technology can help detect potential security threats and monitor suspicious activity in real-time. In addition, in the industrial sector, image recognition is also used to improve efficiency, for example in product quality control on manufacturing production lines, where machines are able to detect defects in products faster and more accurately than human inspection. As such, these increasingly sophisticated image recognition capabilities make a major contribution to effectiveness and security in various sectors (Taye, 2023).

Speech recognition is also very important because it allows for the development of devices capable of understanding and responding to voice commands, creating a more natural and efficient interaction between users and technology. This technology has been implemented in virtual voice assistants such as Siri, Google Assistant, and Alexa, making it easier for users to access information and control smart home devices with just their voice (Jauro et al., 2020). In the healthcare field, voice recognition helps in automatic transcription services for medical records and real-time patient monitoring, thereby reducing the workload of medical personnel. In education, speech recognition allows for interactive language learning, where students can practice pronunciation and hear live feedback. With its role in improving convenience, productivity, and effectiveness in various sectors, voice recognition is becoming an increasingly important technology in the digital era that prioritizes quick and easy interactions (Deng & Li, 2013).

Advances in deep learning models play a key role in accelerating and refining image and voice recognition, allowing this technology to achieve levels of accuracy and reliability that were previously difficult to achieve. In image recognition, deep learning models such as convolutional neural networks (CNNs) have improved the system's ability to recognize and analyze images more deeply, from object classification to the detection of complex features, such as faces and expressions (Bhangale & Kothandaraman, 2022). With these capabilities, CNN-based systems have helped create advanced applications that are now widely used in security, retail, and healthcare. For example, CNNs in security systems can not only recognize faces but also identify suspicious behavior based on movement patterns, while in the medical world, these models are able to detect signs of disease in medical images with a high degree of accuracy (Lionakis et al., 2023).

Similar advances are also happening in the field of speech recognition, where deep learning models such as recurrent neural networks (RNNs) and transformers have brought about a huge leap in understanding and processing voice signals. These models allow devices to recognize complex speech patterns, including intonation, dialect, and context (Akkus et al., 2017), thereby enriching the capabilities of voice assistants and other speech recognition systems. For example, RNNs and transformers make smart devices able to respond to voice commands more accurately and quickly, and understand conversations more naturally. These two models not only improve the quality of human-machine interaction but also expand application opportunities in various sectors, such as education, customer service, and assistant devices for people with disabilities. Thus, advances in deep learning models continue to be the main driver in optimizing the potential of image and voice recognition technology in various aspects of human life (Sarker, 2021; Delić et al., 2019).

Previous research has shown that deep learning technology has changed the landscape of image and voice recognition, with models being further refined to achieve high accuracy. One of the important research in the field of image recognition is the development of Convolutional Neural Networks (CNNs) by Krizhevsky et al. (2012) through the AlexNet model, which brought a major breakthrough in image classification by significantly reducing the error rate in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC). Since then, CNN models have continued to evolve, resulting in more complex variants such as ResNet and EfficientNet, which are capable of handling a wide variety of image recognition tasks, from object detection to image segmentation. This research proves that deep learning models can process complex features in images more effectively than traditional approaches, thus opening up new opportunities for applications in the security, transportation, and health sectors.

Research by Graves et al. (2013) introduced RNN with Long Short-Term Memory (LSTM) as a solution to process long sequence of voice data, thereby improving the speech recognition ability of this model. Further, the emergence of transformers-based models such as BERT and Wav2Vec optimizes speech recognition better in understanding sentence context and improves accuracy in various languages and dialects. The research confirms that deep learning models not only improve accuracy in speech recognition but also enable the development of systems that are more adaptive to user needs, such as intelligent voice assistants and customer service applications. These findings show that deep learning models play an important role in creating more sophisticated image and voice recognition systems that are responsive to user dynamics. Based on this, this study aims to examine the latest developments and future directions of deep learning techniques in image and sound recognition.

2. RESEARCH METHODS

This study uses a systematic literature review method to analyze the latest development and future direction of deep learning techniques in image and sound recognition. The main data sources are scientific articles, conferences, and the latest technical reports relevant in the fields of deep learning, image recognition, and speech recognition, drawn from academic databases such as IEEE Xplore, ScienceDirect, and arXiv. The literature selection process involves filtering by year of publication, focusing on research in the last five years to capture the latest trends and technologies. The inclusion criteria also consider studies that use popular models such as CNNs, RNNs, and transformers, as well as current optimization techniques such as transfer learning and data augmentation. The analysis is carried out by identifying patterns of findings, recent innovations, challenges, and potential improvements in each model and technique used.

In addition to analyzing current trends, this study also applies trend forecasting methods to examine the future direction of deep learning development in image and sound recognition. Using the results of literature analysis, this study projects the potential for new applications and technological developments that enable deep learning to be more efficient and resource-saving, such as through the multimodal model approach and the development of lightweight models. Discussions related to the application of technology in various sectors, such as health, safety, and automotive, were presented to illustrate how these innovations can pave the way for new applications in the future. As such, this methodology not only provides a comprehensive overview of the latest trends but also offers a forward-looking perspective on the opportunities and challenges in the development of deep learning technology.

3. RESULT AND DISCUSSION

Deep Learning

Deep learning is a subfield of machine learning that focuses on algorithms inspired by the structure and function of the human brain, specifically artificial neural networks (ANNs). These networks consist of multiple layers of nodes, or "neurons," which process data in a hierarchical manner (Khanam et al., 2022). The primary feature of deep learning is its ability to automatically learn from large amounts of data by identifying patterns without requiring explicit programming for each task. This self-learning capability allows deep learning models to excel in tasks such as image recognition, speech recognition, natural language processing, and even complex decision-making processes. Over the years, deep learning has evolved significantly, becoming a driving force behind many of the most impactful AI advancements in recent history (Khalil et al., 2019).

One of the key characteristics of deep learning models is their ability to handle unstructured data. Unlike traditional machine learning models that require feature extraction or manual intervention, deep learning models can work directly with raw data such as images, audio, and text (Cummins et al., 2018); Delić et al., 2019). For example, in image recognition, deep learning models, particularly Convolutional Neural Networks (CNNs), automatically learn to identify important features such as edges, textures, and objects through layers of neurons. In speech recognition, Recurrent Neural Networks (RNNs), especially Long Short-Term Memory (LSTM) networks, are used to model sequential data and capture the temporal dependencies in speech. This ability to process unstructured data makes deep learning a versatile and powerful tool across a wide range of domains.

Neural networks, the foundation of deep learning, have gained attention for their deep architecture, where each layer builds on the output of previous layers. This multi-layered approach allows for hierarchical feature learning, where lower layers capture basic features like edges and textures in images, while higher layers recognize more complex patterns, such as faces or objects (Mehrish et al., 2023). This is particularly evident in CNNs, which have set new benchmarks in the field of computer vision. By using multiple layers of convolutional and pooling operations, CNNs have enabled breakthroughs in tasks such as object detection, facial recognition, and autonomous driving. As the number of layers increases, the model's ability to capture intricate patterns also improves, although this also increases the computational demands (Z. Zhang et al., 2018).

A major advancement in deep learning is the use of pre-trained models and transfer learning. These techniques allow deep learning models to be trained on large, general datasets and then fine-tuned for specific tasks, significantly reducing the amount of data and computational resources required for new tasks. For example, models like ResNet and VGGNet have been pre-trained on massive datasets like ImageNet, which contains millions of labeled images (Taye, 2023). These pre-trained models can then be adapted for specific tasks such as medical image analysis, where labeled medical images are scarce. Transfer learning not only accelerates model development but also makes deep learning accessible to smaller organizations with limited resources, democratizing access to cutting-edge AI technology.

Despite its successes, deep learning faces several challenges that hinder its broader adoption and effectiveness. One of the most significant issues is the requirement for large datasets to train models effectively. While deep learning models have shown impressive results when provided with vast amounts of data, they tend to perform poorly when data is limited (Cummins et al., 2018). Moreover, obtaining high-quality, labeled datasets can be expensive and time-consuming, particularly in specialized fields like medical imaging. Additionally, deep learning models are often considered "black boxes" because of their lack of interpretability. This is a concern in fields where transparency and accountability are critical, such as in healthcare and finance. Researchers are actively exploring methods to make deep learning

models more interpretable and explainable, but this remains a major hurdle in the technology's adoption in high-stakes environments (Khanam et al., 2022).

Another challenge associated with deep learning is its computational complexity. Training deep learning models requires significant computational power, often necessitating the use of high-end Graphics Processing Units (GPUs) or specialized hardware like Tensor Processing Units (TPUs). This requirement can make deep learning prohibitively expensive for smaller businesses and researchers without access to such resources. Additionally, the large size of deep learning models can make them slow to deploy and difficult to scale in real-time applications (Jauro et al., 2020). Research is ongoing into more efficient training techniques and lighter models that can be deployed in resource-constrained environments, such as mobile devices or edge computing platforms.

Finally, ethical and societal implications of deep learning cannot be ignored. As deep learning technology becomes more embedded in systems that impact people's lives, such as autonomous vehicles, facial recognition, and hiring algorithms, the potential for misuse and bias grows. If deep learning models are trained on biased data, they can perpetuate harmful stereotypes and make discriminatory decisions (Taye, 2023). Ensuring fairness, transparency, and accountability in deep learning models is crucial, especially when they are used in sensitive applications like criminal justice or hiring practices. Additionally, the widespread use of deep learning could have profound impacts on employment, as automation powered by AI could replace jobs in various sectors. Balancing the benefits of deep learning with these ethical considerations is an ongoing challenge for researchers, policymakers, and businesses alike.

In conclusion, deep learning has revolutionized many areas of artificial intelligence, enabling breakthroughs in image and speech recognition, natural language processing, and more. Its ability to handle complex, unstructured data has made it an indispensable tool in industries ranging from healthcare to entertainment (Delić et al., 2019). However, the challenges of large dataset requirements, computational demands, interpretability, and ethical concerns must be addressed to fully realize the potential of deep learning. With ongoing advancements in research and development, deep learning continues to evolve, offering exciting possibilities for the future of AI and its integration into everyday life (Deng & Li, 2013).

Model CNN (Convolutional Neural Networks)

CNN (Convolutional Neural Networks) models have become one of the most popular and influential deep learning architectures in the field of image recognition. CNNs are specifically designed to address challenges in processing visual data, such as images and videos, in a more efficient way than previous techniques (Acharya et al., 2017). The CNN architecture consists of several main layers, namely the convolution, pooling, and fully connected layers, which allows this model to extract visual features automatically and hierarchically. In the convolution layer, a filter or kernel is used to detect patterns such as lines, textures, and shapes in the imagery. This process allows the model to learn to recognize basic features, which are then combined to detect more complex objects or patterns in the next layers (Q. Zhang et al., 2019).

One of the main advantages of CNNs is their ability to reduce the number of parameters required compared to traditional neural networks. In conventional neural networks, each neuron is connected to all the neurons in the next layer, which can generate a very large number of parameters, especially when processing high-resolution images. However, in CNNs, connections between neurons occur only in certain areas (known as receptive fields), which reduces the number of parameters and makes the model more efficient. In addition, CNN also implemented a weight-sharing technique, where the same filter is used across the image, which

allows the model to recognize the same pattern across different locations of the image (Zhu & Bain, 2017).

The development of CNNs in recent years has resulted in increasingly complex and efficient architectures, such as AlexNet, VGGNet, ResNet, and EfficientNet. AlexNet, introduced by Krizhevsky et al. in 2012, was one of the first models to achieve outstanding results in the ImageNet competition, using multiple layers of deep convolution and pooling. Since then, many innovations in CNN design, such as increased network depth, the use of skip connections in ResNet, and the depthwise separable convolutions technique adopted by EfficientNet to improve computing efficiency without compromising accuracy (Xiao et al., 2019). In addition, transfer learning techniques have become one of the most commonly used approaches in CNN implementations for image recognition. In transfer learning, models that have been trained on large datasets, such as ImageNet, are used as the basis for training models on smaller, more specific datasets. This technique allows the use of pre-trained CNN models to save training time and improve accuracy, especially in cases where available data is limited. This is particularly useful in a variety of applications, such as facial recognition, object detection, and image-based medical diagnosis, where large and diverse datasets may not always be available (Wu, 2017).

The use of CNNs is not limited to image recognition alone, but it has also expanded to other fields such as video processing, text-in-picture recognition, and even medical data analysis. In the medical field, CNNs are used to analyze medical images such as radiology, MRI, and CT scans to detect diseases or abnormalities more accurately and quickly. Research shows that CNNs can produce more accurate detection compared to human medical experts in some cases, such as breast cancer detection on mammography or brain tumor detection on MRI. This shows the great potential of CNNs in improving the quality and efficiency of medical diagnoses (Praveen Chakravarthy et al., 2022).

While CNNs offer many benefits, there are some challenges that must be faced in using them. One of the main challenges is the need for large and diverse training data, which can limit its application in domains that have data limitations. In addition, although CNNs reduce the number of parameters through weight sharing and receptive field techniques, models with very large depths still require high computing power, which can be an obstacle in implementation on devices with limited resources (Sekaran et al., 2020). Therefore, research continues to be carried out to develop more efficient CNN models, such as by using model compression and processing techniques in edge devices to solve this problem. Overall, CNN models have revolutionized the field of image and visual recognition in a highly effective and efficient way. CNN's ability to automatically learn to extract features from images, reduce the number of parameters required, and achieve high accuracy has made it a top choice in a wide range of applications, from facial recognition to medical applications (Kattenborn et al., 2021). With the continuous development of new technologies and architectures, as well as the development of transfer learning techniques, CNN is expected to continue to be the cornerstone for innovation in the field of image recognition and other related fields (Alzubaidi et al., 2021).

Latest Trends in Voice Recognition

Voice recognition technology has seen remarkable advancements in recent years, driven by developments in machine learning, especially deep learning. One of the latest trends in this area is the adoption of deep neural networks (DNNs) and recurrent neural networks (RNNs), particularly Long Short-Term Memory (LSTM) networks, for handling sequential data like speech. These architectures excel at recognizing patterns in speech data over time, making them ideal for voice recognition tasks (Kwon et al., 2018). LSTM networks, which are a variant of RNNs, are capable of handling long-range dependencies in speech signals, addressing the challenge of maintaining context over time. This has significantly improved the accuracy and

robustness of voice recognition systems, especially in noisy environments and with varying accents.

In addition to advancements in model architectures, there has been a surge in the use of transformer-based models for voice recognition. The transformer model, which underpins models like BERT and GPT, has recently been adapted for speech processing tasks such as automatic speech recognition (ASR) and speaker identification (Q. Zhang et al., 2019). Unlike traditional RNNs, transformers rely on self-attention mechanisms that allow them to weigh the importance of different parts of the input sequence, making them highly effective for understanding speech context. This has led to the development of models such as Wav2Vec 2.0, which pre-trains speech representations from raw audio and then fine-tunes them for specific tasks. These transformer-based approaches have demonstrated significant improvements in both accuracy and efficiency for voice recognition systems (Wu, 2017).

Another key trend is the integration of voice recognition with other AI technologies, such as natural language processing (NLP) and computer vision, to create more robust multimodal systems. Voice assistants like Amazon's Alexa, Google Assistant, and Apple's Siri are increasingly able to understand and respond not only to verbal commands but also to the context in which those commands are given (Acharya et al., 2017). For example, these assistants can combine voice inputs with contextual information from users' activities, calendars, and devices to provide more personalized and relevant responses. This trend is pushing the boundaries of voice recognition from simple command execution to more sophisticated, context-aware interactions, which can have significant implications for industries like healthcare, automotive, and customer service.

The rise of voice recognition for multilingual and dialectal understanding is another significant trend. Early voice recognition systems were limited to specific languages or accents, but recent advancements in deep learning have made it possible to develop models that can handle multiple languages and a range of dialects. This progress has been made possible by the availability of large, diverse datasets and the development of cross-lingual models (Fu et al., 2016). For example, models like Google’s Multilingual BERT and Facebook's XLM-R are capable of processing speech in several languages simultaneously, making it easier for global users to interact with voice recognition systems regardless of their native language. As these models improve, voice recognition will become more inclusive and accessible, allowing users from different linguistic backgrounds to benefit from AI-powered voice assistants.

In addition to improving recognition accuracy, another significant trend is the push for privacy-conscious voice recognition systems. With the increasing use of voice assistants in everyday devices, concerns about data security and user privacy have grown. Many companies are now exploring edge computing and on-device processing to ensure that voice data is not sent to centralized servers, thereby reducing the risk of data breaches and unauthorized access. Voice recognition systems that process data locally on devices can offer users more control over their personal information while still providing the benefits of AI. Moreover, techniques like federated learning are being explored (Kwon et al., 2018), where models are trained across multiple devices without sharing sensitive data, ensuring user privacy while benefiting from continuous model improvements.

Finally, the advancement of voice recognition in specialized domains is another area where significant progress is being made. Industries such as healthcare, finance, and legal services are increasingly adopting voice recognition technology to streamline workflows and improve customer experiences. In healthcare, voice recognition is being used for tasks such as medical transcription, where it assists healthcare professionals in documenting patient information efficiently. In finance, it is used for secure authentication, enabling users to access accounts or make transactions with just their voice. Similarly, in legal services, voice

recognition is employed to transcribe court hearings and depositions. As these specialized applications grow, voice recognition technology is expected to become more tailored to the specific needs of each sector, improving its precision and efficiency in these contexts (Q. Zhang et al., 2019)

In summary, the latest trends in voice recognition highlight the tremendous advancements in model architectures, integration with other AI technologies, multilingual capabilities, privacy concerns, and sector-specific applications. With the increasing sophistication of deep learning models, voice recognition technology is becoming more accurate, efficient, and accessible, paving the way for a future where voice interfaces will be an integral part of many industries and daily life (Wu, 2017). These developments not only enhance the user experience but also open new opportunities for innovation in AI-powered voice applications across a range of fields.

Challenges and Obstacles in the Development of Deep Learning Technology for Image and Voice Recognition

The development of deep learning technology for image and voice recognition has seen remarkable progress over the past decade, yet significant challenges and obstacles remain. One of the primary challenges is the requirement for large and diverse datasets (Z. Zhang et al., 2018). Deep learning models, especially convolutional neural networks (CNNs) for image recognition and recurrent neural networks (RNNs) or transformers for voice recognition, perform best when trained on massive amounts of data. However, obtaining high-quality, labeled datasets for both image and voice recognition is often difficult and expensive. For instance, in image recognition, obtaining thousands or millions of labeled images across various categories can be a logistical challenge(Taye, 2023). Similarly, for voice recognition, creating diverse datasets that cover multiple accents, dialects, languages, and speaking styles requires significant time and resources. The lack of sufficiently large and varied datasets can lead to underperforming models that are biased or unable to generalize well across real-world situations (Sarker, 2021).

Another challenge is computational complexity and resource demands. Deep learning models, especially those with many layers and parameters, require enormous computational power for both training and inference (Khalil et al., 2019). For example, training state-of-the-art CNNs for image recognition or transformer models for voice recognition involves processing vast amounts of data through multiple iterations, which demands specialized hardware such as Graphics Processing Units (GPUs) or Tensor Processing Units (TPUs). These resources are often expensive and can be inaccessible for smaller research labs, startups, or businesses with limited budgets. Moreover, as deep learning models grow in complexity, they require more memory and computational power, making them difficult to deploy in real-time applications, particularly in environments with limited hardware, such as smartphones or edge devices (Cummins et al., 2018).

The lack of interpretability and transparency in deep learning models also presents a significant obstacle. While these models have demonstrated impressive performance in tasks like image and voice recognition, they often function as "black boxes," meaning that it is difficult for researchers and developers to understand exactly how they make decisions (Akkus et al., 2017). This lack of transparency can be particularly problematic in applications where accountability and trust are paramount, such as in healthcare, autonomous driving, or law enforcement. For example, if a voice recognition system misinterprets a command, or an image recognition model makes an error in medical imaging, understanding the underlying causes of the mistake is crucial to improving the system and ensuring its reliability. Researchers are

actively working on techniques like explainable AI (XAI) to address these issues, but interpretability remains a challenge for many deep learning applications.

Bias and fairness also present significant challenges in deep learning for image and voice recognition. Deep learning models are trained on historical data, and if this data reflects biases present in society, these biases can be inadvertently learned and perpetuated by the model. For example, image recognition systems have historically struggled with underperforming on images of people from diverse racial or ethnic backgrounds because the datasets they were trained on were not representative of this diversity (Dargan et al., 2020). Similarly, voice recognition systems may perform poorly on certain accents or dialects, leading to inequality in user experiences. Addressing these biases requires careful curation of training data, as well as the development of algorithms that can detect and mitigate bias, ensuring that deep learning technologies are equitable and fair for all users.

The generalization and robustness of deep learning models in real-world environments is another significant challenge. While deep learning models often achieve high accuracy in controlled, clean environments, their performance can degrade dramatically when exposed to noisy, dynamic, or adversarial conditions. For example, in image recognition, changes in lighting, angles, or occlusions can negatively affect model performance. Similarly, in voice recognition, background noise, different microphones, or variations in speech rate and intonation can cause the system to fail (Khanam et al., 2022). Developing models that can generalize well to a wide range of conditions, while remaining robust against these challenges, is a key area of ongoing research. This requires both more sophisticated model architectures and techniques such as data augmentation and adversarial training to make models more resilient to real-world variability.

Finally, privacy and security concerns are a growing issue in the development of deep learning technologies for image and voice recognition. As these technologies become more widely used in personal devices, home assistants, and surveillance systems, concerns about how personal data is collected, stored, and used have intensified (Bhangale & Kothandaraman, 2022). In voice recognition systems, for instance, there is the potential for sensitive information, such as private conversations, to be recorded and misused if not properly handled. Similarly, image recognition systems deployed in public spaces can raise concerns about mass surveillance and the invasion of privacy. Ensuring that deep learning models respect user privacy and comply with regulations such as GDPR is crucial for their widespread adoption. Techniques like federated learning, which allows models to be trained on local devices without transmitting personal data to central servers, are being explored to address these concerns and enhance privacy protections (Lionakis et al., 2023).

In conclusion, while deep learning technology for image and voice recognition holds immense potential, it faces several challenges that need to be addressed to fully realize its capabilities (Akkus et al., 2017). These include the need for large, diverse datasets, significant computational resources, model interpretability, fairness and bias mitigation, robustness in real-world conditions, and privacy concerns. Overcoming these obstacles requires continued research and innovation across multiple areas, including data collection, model architecture, and ethical considerations. As these challenges are addressed, deep learning for image and voice recognition is expected to evolve further, enabling more powerful and reliable systems across a wide range of applications (Khalil et al., 2019).

Conclusion

From the results of this study, it can be concluded that the advancement of deep learning techniques, especially in the field of image and voice recognition, has undergone significant

development, with the application of the latest models such as Convolutional Neural Networks (CNNs) for image recognition and Transformer-based models for speech recognition. In image recognition, CNN models continue to show excellent performance in object classification and image detection, thanks to their ability to extract complex features through convolutional layers. Meanwhile, in speech recognition, the use of Transformer-based models, such as Wav2Vec 2.0, has provided a tremendous increase in accuracy in speech recognition automation, thanks to its ability to understand a broader context in the sequence of voice data. Both technologies have proven effective in dealing with challenges such as data variation, different accents, and noise, although challenges related to dataset diversity and computing resource requirements remain significant constraints.

However, the study also identifies various key challenges that must be faced in the further development of deep learning technology for image and voice recognition. One of the biggest challenges is the need for larger and more diverse datasets to reduce bias in models as well as improve accuracy, especially in more varied real-world environments. In addition, although deep learning models have achieved excellent performance, issues related to interpretability and transparency in decision-making are still a major obstacle, especially in applications that require a high level of accountability, such as in the field of health and safety. Going forward, it is important to develop solutions that can address these issues, while ensuring that these technologies remain effective and accessible to more users in various application contexts.

REFERENCE

Acharya, U. R., Oh, S. L., Hagiwara, Y., Tan, J. H., Adam, M., Gertych, A., & Tan, R. S. (2017). A deep convolutional neural network model to classify heartbeats. *Computers in Biology and Medicine*, *89*, 389–396. https://doi.org/10.1016/j.compbiomed.2017.08.022

Akkus, Z., Galimzianova, A., Hoogi, A., Rubin, D. L., & Erickson, B. J. (2017). Deep Learning for Brain MRI Segmentation: State of the Art and Future Directions. *Journal of Digital Imaging*, *30*(4), 449–459. https://doi.org/10.1007/s10278-017-9983-4

Al-Fraihat, D., Sharrab, Y., Alzyoud, F., Qahmash, A., Tarawneh, M., & Maaita, A. (2024). Speech Recognition Utilizing Deep Learning: A Systematic Review of the Latest Developments. *Human-Centric Computing and Information Sciences*, *14*(March). https://doi.org/10.22967/HCIS.2024.14.015

Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M. A., Al-Amidie, M., & Farhan, L. (2021). Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. In *Journal of Big Data* (Vol. 8, Issue 1). Springer International Publishing. https://doi.org/10.1186/s40537-021-00444-8

Bhangale, K. B., & Kothandaraman, M. (2022). Survey of Deep Learning Paradigms for Speech Processing. In *Wireless Personal Communications* (Vol. 125, Issue 2). https://doi.org/10.1007/s11277-022-09640-y

Cummins, N., Baird, A., & Schuller, B. W. (2018). Speech analysis for health: Current state-of-the-art and the increasing impact of deep learning. *Methods*, *151*, 41–54. https://doi.org/10.1016/j.ymeth.2018.07.007

Dargan, S., Kumar, M., Ayyagari, M. R., & Kumar, G. (2020). A Survey of Deep Learning and Its

Applications: A New Paradigm to Machine Learning. *Archives of Computational Methods in Engineering*, *27*(4), 1071–1092. https://doi.org/10.1007/s11831-019-09344-w

Delić, V., Perić, Z., Sečujski, M., Jakovljević, N., Nikolić, J., Mišković, D., Simić, N., Suzić, S., & Delić, T. (2019). Speech technology progress based on new machine learning paradigm. *Computational Intelligence and Neuroscience*, *2019*. https://doi.org/10.1155/2019/4368036

Deng, L., & Li, X. (2013). Machine learning paradigms for speech recognition: An overview. *IEEE Transactions on Audio, Speech and Language Processing*, *21*(5), 1060–1089. https://doi.org/10.1109/TASL.2013.2244083

Fu, S. W., Tsao, Y., & Lu, X. (2016). SNR-aware convolutional neural network modeling for speech enhancement. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, *08-12-September-2016*, 3768–3772. https://doi.org/10.21437/Interspeech.2016-211

Jauro, F., Chiroma, H., Gital, A. Y., Almutairi, M., Abdulhamid, S. M., & Abawajy, J. H. (2020). Deep learning architectures in emerging cloud computing architectures: Recent development, challenges and next research trend. *Applied Soft Computing Journal*, *96*, 1–91. https://doi.org/10.1016/j.asoc.2020.106582

Kattenborn, T., Leitloff, J., Schiefer, F., & Hinz, S. (2021). Review on Convolutional Neural Networks (CNN) in vegetation remote sensing. *ISPRS Journal of Photogrammetry and Remote Sensing*, *173*(November 2020), 24–49. https://doi.org/10.1016/j.isprsjprs.2020.12.010

Khalil, R. A., Jones, E., Babar, M. I., Jan, T., Zafar, M. H., & Alhussain, T. (2019). Speech Emotion Recognition Using Deep Learning Techniques: A Review. *IEEE Access*, *7*, 117327–117345. https://doi.org/10.1109/ACCESS.2019.2936124

Khanam, F., Munmun, F. A., Ritu, N. A., Saha, A. K., & Mridha, M. F. (2022). Text to Speech Synthesis: A Systematic Review, Deep Learning Based Architecture and Future Research Direction. *Journal of Advances in Information Technology*, *13*(5), 398–412. https://doi.org/10.12720/jait.13.5.398-412

Kwon, Y. H., Shin, S. B., & Kim, S. D. (2018). Electroencephalography based fusion two-dimensional (2D)-convolution neural networks (CNN) model for emotion recognition system. *Sensors (Switzerland)*, *18*(5). https://doi.org/10.3390/s18051383

Lionakis, E., Karampidis, K., & Papadourakis, G. (2023). Current Trends, Challenges, and Future Research Directions of Hybrid and Deep Learning Techniques for Motor Imagery Brain–Computer Interface. *Multimodal Technologies and Interaction*, *7*(10). https://doi.org/10.3390/mti7100095

Mehrish, A., Majumder, N., Bharadwaj, R., Mihalcea, R., & Poria, S. (2023). A review of deep learning techniques for speech processing. In *Information Fusion* (Vol. 99). https://doi.org/10.1016/j.inffus.2023.101869

Praveen Chakravarthy, S., Gunasundari, C., Selva Bhuvaneswari, K., Sharma, B., & Chowdhury, S. (2022). Convolutional Neural Network (CNN) for Image Detection and Recognition in Medical Diagnosis. *IET Conference Proceedings*, *2022*(26), 357–361. https://doi.org/10.1049/icp.2023.0579

Sarker, I. H. (2021). Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. *SN Computer Science*, *2*(6), 1–20. https://doi.org/10.1007/s42979-021-00815-1

Sekaran, K., Chandana, P., Krishna, N. M., & Kadry, S. (2020). Deep learning convolutional neural network (CNN) With Gaussian mixture model for predicting pancreatic cancer. *Multimedia Tools and Applications*, *79*(15–16), 10233–10247. https://doi.org/10.1007/s11042-019-7419-5

Taye, M. M. (2023). Understanding of Machine Learning with Deep Learning : *Computers MDPI*, *12*(91), 1–26.

Wu, J. (2017). Introduction to Convolutional Neural Networks. *Introduction to Convolutional Neural Networks*, 1–31. https://web.archive.org/web/20180928011532/https://cs.nju.edu.cn/wujx/teaching/15_CNN.pdf

Xiao, Y., Xing, C., Zhang, T., & Zhao, Z. (2019). An Intrusion Detection Model Based on Feature Reduction and Convolutional Neural Networks. *IEEE Access*, *7*, 42210–42219. https://doi.org/10.1109/ACCESS.2019.2904620

Zhang, Q., Zhang, M., Chen, T., Sun, Z., Ma, Y., & Yu, B. (2019). Recent advances in convolutional neural network acceleration. *Neurocomputing*, *323*, 37–51. https://doi.org/10.1016/j.neucom.2018.09.038

Zhang, Z., Geiger, J., Pohjalainen, J., Mousa, A. E. D., Jin, W., & Schuller, B. (2018). Deep learning for environmentally robust speech recognition: An overview of recent developments. *ACM Transactions on Intelligent Systems and Technology*, *9*(5), 1–16. https://doi.org/10.1145/3178115

Zhu, X., & Bain, M. (2017). *B-CNN: Branch Convolutional Neural Network for Hierarchical Classification*. http://arxiv.org/abs/1709.09890