SHS Web of Conferences **197**, 03002 (2024) https://doi.org/10.1051/shsconf/202419703002
ICADE 2023

Digital audio preservation for Indonesian traditional vocal recognition based on machine learning: A literature review and bibliometric analysis

Hafizhah Insani Midyanti[1*], *Yudi* Sukmayadi[2], *Fensy* Sella[2] and *Dwi Marisa* Midyanti[3]

[1]Program Studi Musik, Universitas Pendidikan Indonesia, Bandung, Indonesia
[2]Program Studi Pendidikan Seni Musik, Universitas Pendidikan Indonesia, Bandung, Indonesia
[3]Program Studi Sistem Komputer, Universitas Tanjungpura, Pontianak, Indonesia

Abstract. The study aims to save Indonesia's extensive voice history by comprehensively examining existing literature and doing a bibliometric analysis. This approach provides a comprehensive understanding of this field's development, methodology, obstacles, and potential future paths. The key focus is machine learning approaches to identify and safeguard Indonesian traditional vocals using several methods, like spectrogram-based techniques, convolutional and recurrent neural networks, transfer learning, attention mechanisms, and hybrid learning. Examining these technologies considers Indonesia's voice variety, providing insights into their adaptability to handling distinct scales, tunings, and stylistic variances. The study incorporates a bibliometric analysis to measure the expansion of literature and ascertain the prominent authors, journals, and keywords in this developing topic. This study improves our comprehension of the research terrain and the conceptual paths that drive the progress of the field. Indonesia's traditional vocal music faces the imminent challenges of industrialization and globalization. However, there is hope for developing machine learning to preserve digital audio data of traditional music, especially traditional vocals in Indonesia, some of which are almost extinct. We explore the use of machine learning to honour and protect Indonesia's varied vocal traditions while also considering the ethical responsibilities associated with this undertaking.

1 Introduction

Recently, substantial advancement in audio machine learning has resulted in profound alterations in our understanding and engagement with auditory stimuli [1,2]. This technology has been employed in many domains, spanning speech recognition, music analysis, and other disciplines [3,4]. Recognizing and preserving traditional vocal

* Corresponding author: dicemidyanti@upi.edu

practices hold significant cultural and aesthetic significance within this ever-changing context [5].

The recognition of Indonesian traditional vocal performances serves as a starting point for a musical exploration of the diverse and broad musical heritage found within the archipelagic nation of Indonesia. This cultural heritage encompasses traditional vocal styles in various geographical regions and ethnic populations. The diversity of aural expressions that characterize the vocal repertoire in Indonesia, including the magical melodies of Beluk Sundanese singing [6,7], energetic vocalizations seen in Balinese kecak chants [8,9] and many other singing cultures that exist in Indonesia. Indonesia's vocal traditions provide evidence of the profound cultural exchanges and creative progress that have transpired over generations [10].

The vocal styles traditionally used in Indonesia significantly correlate with various ceremonial acts, rituals, and everyday life, offering a unique perspective into the people's profound emotions and fundamental nature [11,12]. Nevertheless, significant challenges must be overcome despite the historical importance attributed to the conservation and recognition of these indigenous traditional vocals. Various efforts to maintain a society's identity and cultural richness are closely related to the preservation of traditional music and its cultural heritage, one of which is traditional vocals, which have high cultural value and therefore need preservation efforts. The existence of several traditional vocals is increasingly being abandoned due to developments over time, so efforts are needed to preserve and maintain them so that their existence is remembered in society.

Digital audio preservation is a new approach to digital technology to record, store, and protect traditional music. Digital preservation is defined as a specific act of technology. The underlying, long-term goal is to ensure immortality and access to digital documents, regardless of support, format, or system [12,13]. Formats, resolution levels, support, and technological systems that meet international standards are necessary for sound preservation [14]. Versions outside these standards are not secure in the long term for migration and information exchange or are associated with emerging formats [15]. In addition, the fusion of vocal and instrumental elements in Indonesian traditional music creates a captivating and complex auditory environment that merits scientific investigation [16,17].

It is crucial to recognize the significance of audio machine learning in addressing difficulties in preserving digital audio. The utilization of machine learning techniques and developments in audio signal processing has opened up new possibilities for identifying and analyzing complex speech patterns [18,19] in the context of Indonesian traditional music [20,21]. These approaches enable researchers and practitioners to effectively analyze vocal styles while enhancing their understanding of the underlying structural and cultural contexts connected with them.

In the contemporary period characterized by rapid technological progress, there is a pressing need to proficiently employ machine learning methodologies to document, preserve, and disseminate Indonesia's rich and varied vocal traditions. The present literature review thoroughly examines the current corpus of research on this particular topic, shedding light on essential methodologies, sources of data, challenges, and prospective directions for future exploration. Participating in this activity significantly contributes to the ongoing scholarly discussion on the intersection of technology and culture, emphasizing the role of machine learning in bridging modern innovations and traditional practices.

The study is on the methodology of obtaining and preparing audio data for the goal of machine learning, with specific emphasis on the unique challenges experienced while working with Indonesian vocal recordings. Machine learning involves examining diverse approaches and algorithms that empower computers to acquire knowledge and generate predictions or conclusions autonomously without explicit programming. This comprehensive literature review aims to examine the current state of audio machine learning in the specific domain of identifying traditional vocal styles. This research aims to provide a valuable resource for researchers, artists, and enthusiasts in this field regarding a journey that celebrates cultural history, showcases technological advancements, and ultimately deepens our understanding of the significant impact of technology in preserving and progressing traditional arts.

2 Method

This study involves a thorough analysis of the existing research in the field, utilizing bibliometric techniques in audio machine learning for Indonesian traditional vocal recognition by exploring trends, theories, literature reviews, and bibliometric analysis. We performed a literature review and bibliometric analysis using VOSviewer that included studies published for five years (2019-2023) regarding integrating the audio machine learning for Indonesian traditional vocal recognition.

3 Results and discussion

3.1 Digital audio preservation

Digitalization and digital preservation are intricately interconnected ideas. Voutssas Marquez states that digital preservation is a distinct technological process. The ultimate objective is to guarantee the perpetuity and availability of digital documents, irrespective of their support, format, or system. It is necessary to undertake the same maintenance level as required for the permanent protection and security of the data. When material or data sustains damage, it is imperative to endeavor to restore it.

Térmens [22] posits that digital preservation guarantees the future accessibility and utilization of digital documents generated in the present or past through information conservation and security, ensuring its long-term maintenance and usage. According to Candás-Romero [23], digital preservation is a continuous and intricate worldwide undertaking that considers information's physical and logical aspects. It involves using suitable, sufficient, and formal document descriptions. Therefore, digital preservation of sound documents is an enduring approach to safeguarding, organizing, and managing digital audio content (also known as essence or media), facilitating its long-term accessibility, distribution, reuse, and accompanying metadata.

Conversely, digitalization refers to converting content stored in analogue to digital format. It involves replacing analogue signals with digital ones. Digitization safeguards precious records by preventing tampering and preserving them from harm. Ensuring the survival of audiovisual content is only possible through this method [24]. Digitalization serves as both an analogy for transferring to digital platforms and converting processes and a method in file management [25]. By use of the digitization process, specific content can be transformed. As an illustration, an audio recording is transformed into a sequence of numerical values. According to the recommendations of experts and researchers [13], digitalization is based on the following recommendations:

- Digitization should be performed without applying data compression, adhering to the agreed-upon quantification and sampling of digital signals.
- It is necessary to accurately capture analogue signals from documents using suitable recording and reproduction equipment.
- The digitization process should be conducted while preserving the source material without any alterations.
- An alphanumeric code must be provided to establish a connection between the media and the metadata to identify the material.

Digitization is the process of converting analogue documents into a digital format in order to preserve them [26]. This process is carried out using digital preservation techniques. Aside from providing definitions for digitization and digital preservation, it is crucial to remember that many documents are generated regularly using digital means. The complexity of arguments around digital preservation arises from the need to address digital content and content first captured using digital technology [27]. Digital conservation encompasses a range of interventions and actions aimed at preserving and safeguarding digital resources. Transitioning from obsolete assistance to contemporary support is a method of conservation. Preventive measures aim to mitigate harm but do not guarantee the avoidance of obsolescence. Wright's definition states that migration is a preservation component, and digital document conservation is continuous digital preservation.

Chen predicted over ten years ago that conservation and preservation will undergo a transformation and become digital preservation. The improvements have significantly altered the security of sound archives by introducing new professional profiles or genres, creating lossless digital copies, and integrating media and metadata as fundamental elements of digital preservation. These modifications necessitate the implementation of digital storage and management systems for large amounts of data and the enhancement of accessibility, distribution, and utilization of audio files. Furthermore, this transition encounters many obstacles stemming from the digital conservation of audio records.

Incorporating digitization marked the initial addition to the conventional practice of handling sound files. According to Westerhof, digitalization is the fundamental cause of organizational changes and the development of issues with files. This process also impacts our cognitive perception of archiving. Digitalization has created new processes for identifying files and entering digital documents. This workflow encompasses the validation and verification of cataloging and digitization and addressing matters related to remote access [28,29]. It also involves continuously ensuring the integrity and consistency of audio signals and metadata and creating periodic copies of media and metadata in compliance with archive policies [30].

Due to the rise of digitalization, analogue copies are now rarely produced. Using digital platforms to secure a copy and guarantee long-term preservation is advisable.

Lossless copying can only be achieved in the digital realm. Hence, digital preservation's primary focus is safeguarding digital content in its original digital format.

As previously mentioned, the fundamental elements to be considered in digital preservation are digital sound, also known as essence or media, and metadata. In order to preserve sound, it is crucial to have formats, resolution levels, support, and technological systems that adhere to international standards. Versions must adhere to these standards to be considered secure in the long run for migration and information interchange. They may also be linked to new formats [24]. Metadata is essential data for utilizing and administrating sound collections following a digital transformation. Metadata serves the purpose of identifying and organizing information and facilitating its retrieval [31]. De Jong stated that digital sound file metadata is created from the information obtained via classifying, digitizing, and managing sound files. An effectively designed digital archive will automatically produce metadata and incorporate details about the original recording medium, format, and preservation condition, as well as the playback equipment and standards, digital resolution, all utilized equipment, operators and performers, and any other involved processes or procedures [24].

Sound file metadata serves as the primary communication between existing and upcoming technology systems [32]. De Jong stated that devoid of metadata, digital information interchange becomes unattainable. Sound files must comprehensively record all modifications made to the metadata. When metadata registries are built and maintained correctly, they can be safeguarded against modifications or the inclusion of pertinent material in suitable fields [33].

In recent years, the expansion of sound and audiovisual archive collections has presented significant obstacles regarding storage capacity. Wright defines mass digital storage and management systems as comprehensive, automated systems that store, manage, maintain, distribute, and preserve digital objects and their associated metadata, backup systems, and primary storage. Bulk digital management and storage systems streamline and automate regulating, digitizing, storing, classifying, administering, and distributing digital objects and metadata of sound files. The primary goal is to ensure the preservation and accessibility of these digital assets.

The most significant changes from digital preservation are enhancing access, visibility, and the ability to reuse sound files. According to Thibodeau, the significance of digital preservation lies in the degree to which the information is utilized. Hence, the primary objective of conservation is to maximize the potential for exploitation. The utilization of digital sound files is enhanced by internet connectivity and social media platforms, leading to the development of novel specialized services that rely on sound data catalogues. These catalogues are vital in evaluating the contents of each archive.

Digital preservation signifies a fundamental shift in advancing the culture of listeners [34]. From this standpoint, tasks such as selecting audio libraries suitable for preschool children, constructing sound maps, or designing educational or cultural content maps are only a few instances of the diverse possibilities offered by reusing sound files. A sound file possesses energy, and the accumulated recordings serve as a living heritage and an essential element of modern civilization.

The task can be defined as the endeavor carried out by digitally preserving sound files to guarantee long-term digital preservation. According to Thibodeau, among the issues faced, obsolescence and migration are particularly noteworthy. It is stated that the only certainty in information technology is the constant evolution of technology. Consequently, digital preservation emerged as a long-term strategy to save audio archives, addressing the deterioration or becoming outdated analogue recordings made in earlier years that may have been inactive for a minimal period. With the progression of technology and the growth in storage capacity, hardware frequently needs to be updated.

Similarly, as computer processors and various software applications become old, users can no longer view numerous files that rely on these components. Hence, any determination concerning digital preservation necessitates the incorporation of technological adaptability. Bradley asserts that a durable and enduring vision for digitally preserving sound files is essential. Consequently, migration will continue to occur often. Schuller and Teruggi also highlight the significance of transferring information to new storage systems due to the imminent obsolescence of older storage systems as a crucial aspect of digital preservation. As per Schuller, migration is a recurring process that should be completed on time for a decade.

3.2 Machine learning

Arthur Samuel introduced the concept of machine learning in 1959. During that period, machine learning research mainly focused on basic statistical models within the domain of computer games [35]. Tom M. Mitchell [36] offered a refined definition of machine learning, emphasizing the automatic enhancement of mathematical models' performance through experience. Several machine learning techniques have been developed and commonly employed, including backpropagation neural networks, which were created in the 1970s [37], support vector machines, which were developed in the 1990s [38], and deep learning, which emerged in the 2000s [39]. Machine learning encompasses three primary paradigms, as Jordan and Mitchell [40] outlined: supervised learning, unsupervised learning, and reinforcement learning. Every data sample is consistently associated with a specific label in supervised learning. A mapping model receives data samples as input and aims to learn output predictions that closely match the relevant labels. It implies that the labels "supervised" the training process [41].

Unlike supervised learning, unsupervised learning does not involve any labeling for data samples. Mapping models aim to reveal inherent resemblances in data samples [42]. Reinforcement learning occupies a position between supervised and unsupervised learning, as it involves data samples that lack explicit labeling. Instead, a reward is assigned as a measurement for each Action. The mapping model seeks to identify the most favorable sequence of actions in order to maximize the total rewards [43]. In addition to the three primary machine learning paradigms, various sub-paradigms exist. One such sub-paradigm is semi-supervised learning, which combines elements of both supervised and unsupervised learning paradigms. Unsupervised learning uses unlabeled data samples to enhance supervised learning, as obtaining unlabeled data samples is comparatively easier than obtaining labeled data samples. Two often employed paradigms in audio machine learning are supervised learning and unsupervised learning.

3.2.1 Supervised learning

Supervised learning is the dominant approach in machine learning, as stated by Jordan and Mitchell [40]. Many current applications heavily rely on supervised learning, such as sound event detection [44-46], speech recognition [47-49], action recognition [50,51], and machine translation [52-54]. Several supervised learning methods have been created, including deep learning [55], decision trees [56], nearest neighbors [57], and support vector machines [58].

The advancements in deep learning have been remarkable in recent years, driven mainly by the surge in data volumes and the reduction in hardware expenses. The technique described by Lecun et al. has gained significant prominence in several applications. Deep learning is a specialized area within machine learning that uses numerous processing layers to create computational models. Several deep learning architectures have been proposed, each with different designs for the processing layers. These include Deep Belief Networks, Deep Auto-Encoder [59], Deep Stack Networks [60], Convolutional Neural Networks (CNNs) [61-63], and Recurrent Neural Networks [64-66].

Of all the deep learning architectures, Recurrent Neural Network (RNN) has been demonstrated to be efficient in handling sequential data and recognizing long-term connections stated by Lecun et al., because RNNs possess specifically designed iterative operations in their hidden layers, enabling them to transmit historical information from previous inputs and hidden layers to subsequent ones. Nevertheless, in multi-class classification tasks, most existing RNNs employ cross-entropy loss functions [55]. However, these loss functions only partially use the information conveyed by data labels because the cross-entropy loss function solely focuses on the target category and does not consider the competing categories during the training process [67]. Junbo Ma devised a Large-Margin Recurrent Neural Network (LMRNN) that incorporates the large-margin discriminative principle as a heuristic term to guide the convergence process in training.

3.2.2 Unsupervised learning

Unsupervised learning is independent of data labels, unlike supervised learning. Obtaining or accessing data labels might be challenging in real-world scenarios. There are instances where it is necessary to examine the fundamental framework of the data sample [40]. Unsupervised learning is relied upon by various applications, such as clustering [68-70], dimensionality reduction [71-73], feature extraction [74-77], and representation learning [78,79].

Various unsupervised learning algorithms, including k-means [80], Gaussian mixture model [81,82], Auto-encoder [83], and Generative Adversarial Networks (GANs) [84,85], have been developed. Clustering is a fundamental technique in statistical data analysis that aims to categorize data samples into subgroups using a predefined set of criteria [80,86].

Several clustering algorithms, such as k-means [87], k-medoids [88], hierarchical clustering [89], and density-based clustering [90], have been developed. Multi-view clustering has gained popularity in recent years because of the prevalence of data samples that consist of many views, each providing distinct viewpoints on the data sample [91,92]. Multi-view clustering aims to enhance the accuracy and generalization of clustering by leveraging complementary information from several perspectives [93]. Some research employs subspace techniques to analyze a shared latent subspace across different perspectives and subsequently conduct clustering [93,94]. However, most existing multi-view subspace clustering techniques rely on either k-means or spectral clustering algorithms. In these methods, the number of clusters (k) and the weights assigned to distinct views must be manually set [95,96].

3.3 Machine learning techniques for audio recognition

Machine learning approaches have become a potent tool in the field of audio recognition for the identification and classification of sound patterns [97,98]. The application of machine learning is particularly significant in maintaining and recognizing Indonesian traditional vocal music, which contains rich cultural and stylistic characteristics in its audio. There are several machine learning methodologies and their suitability for identifying Indonesian traditional vocal patterns.

3.3.1 Spectrogram-based analysis

Spectrogram-based techniques encompass the transformation of audio impulses into visual depictions called spectrograms [99,100]. These visualizations depict the temporal evolution of the frequency content of a sound. Machine learning algorithms can be trained to analyze spectrograms for audio recognition and detect patterns that indicate distinct voice styles [101,102] or traditions. This method is particularly beneficial for identifying vocal embellishments and fluctuations in tone and quality that are distinctive to vocal [103].

3.3.2 Convolutional Neural Networks (CNN)

Convolutional Neural Networks, typically employed in image recognition, have been modified to suit audio recognition problems. Convolutional Neural Networks (CNNs) employ convolutional layers to identify and analyze patterns or distinctive characteristics within the given data [104,105]. These networks can analyze spectrograms or time-frequency representations to identify audio patterns in audio recognition [106,107]. CNNs can be trained to recognize certain vocal attributes of Indonesian traditional music, such as the complex melodies of Sundanese Kawih or the subtle phrasing of Beluk singing.

3.3.3 Recurrent Neural Networks (RNN)

Recurrent Neural Networks (RNNs) are particularly suitable for jobs that involve data arranged in a sequence [108], which makes them highly helpful for audio recognition [109,110]. RNNs can capture the sequential relationships in audio signals within the framework of Indonesian vocal music. They can be instructed to identify specific sounds or tones and the changing arrangements and musical compositions crucial to the heritage.

Recurrent Neural Networks (RNNs) are especially advantageous when dealing with extended voice performances; accurately comprehending the cadence and tempo is paramount.

3.3.4 Long Short-Term Memory (LSTM)

LSTM, short for Long Short-Term Memory, is a specialized Recurrent Neural Network (RNN) that is particularly effective at recording and understanding long-term relationships in sequential data [111,112]. Indonesian traditional vocal music frequently showcases complex melodies and elaborate embellishments that span considerably. LSTMs are adept at capturing these extended temporal patterns, rendering them invaluable for discerning the nuanced vocal expressions seen in Indonesian music. LSTMs have demonstrated their use in identifying prolonged vocal phrases and comprehending the intricacies of performances.

3.3.5 Transfer learning

Transfer learning utilizes pre-trained models built on large datasets to perform audio identification tasks [113-115]. Transfer learning is a helpful method in Indonesian traditional vocal music, especially when labeled data are scarce. By refining pre-trained models using a smaller dataset focusing on Indonesian vocal styles, researchers can achieve more precise recognition while saving computational resources.

3.3.6 Mechanisms of attention

Attention mechanisms play a vital role in comprehending the significant features of audio data. They allow machine learning algorithms to concentrate on the most pertinent components inside the audio signal [116,117]. This function is valuable for identifying Indonesian traditional singers since it can emphasize distinct melodic elements, embellishments, or distinctive stylistic attributes in the audio.

3.3.7 Hybrid models

In order to tackle the intricacy of Indonesian traditional vocal music, it is possible to utilize hybrid models that include different machine learning techniques. An example of a more thorough approach to audio recognition involves the combination of Convolutional Neural Networks (CNNs) for extracting features from spectrograms and Recurrent Neural Networks (RNNs) for capturing temporal dependencies[118,119].

3.4 Bibliometric analysis of machine learning for Indonesian traditional vocal recognition areas based on VOSviewer using network visualization

Figure 1 shows the network visualization for machine learning and Indonesian traditional vocal recognition areas. Figure 2 shows machine learning and Indonesian traditional vocal recognition areas based on VOSviewer using density visualization.

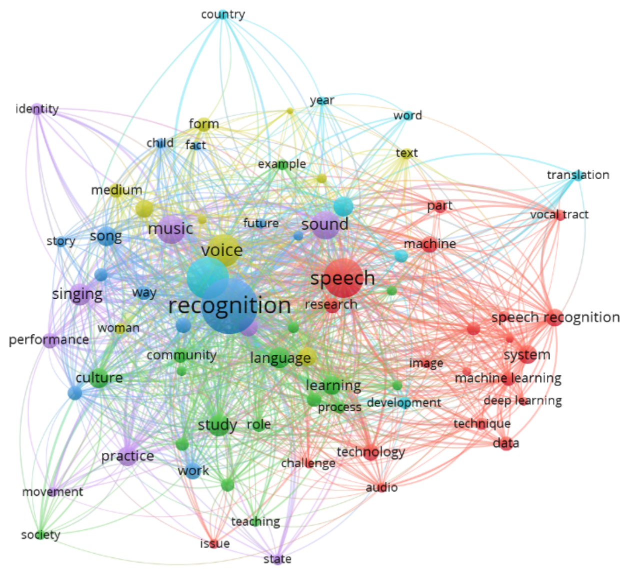

Fig. 1. Network visualization for machine learning and Indonesian traditional vocal recognition areas.

In the past five years, there has been a significant increase in research in machine learning and vocal recognition. It has resulted in a dynamic and diverse environment covering various interconnected topics. This bibliometric analysis has revealed a diverse range of studies that not only investigate the complexities of vocal recognition technology but also examine its significant associations with history, music, sound, language, community, culture, technology, challenges, teaching, the vocal tract, words, texts, years, singing, practice, etc. The historical aspect of this study landscape demonstrates the aspiration to comprehend the development leading to the present state of vocal recognition technology. Researchers have illuminated the journey of vocal recognition from its crude beginnings to the advanced speech-to-text solutions of today by tracing its roots and mapping its evolutionary trajectory. The historical insights offer a crucial context for the evolution of the subject and establish a basis for future advancements.

The convergence of music and vocal recognition epitomizes a seamless amalgamation of art and technology. Music is a fundamental component of human culture and has significant implications for studies on vocal recognition. This field of research aims to improve the process of finding and categorizing music by utilizing advanced machine-learning methods. This study examines sound quality, tone, and linguistic characteristics in music and speech, pushing the limits of vocal recognition systems' capabilities. Language is vital in research, serving as the foundation for vocal recognition. The accuracy and adaptability of vocal recognition systems are improved by rigorously examining linguistic factors, which include phonetics and semantics. The enormous range of languages and dialects around the world requires us to examine the cultural background to understand them. Researchers diligently strive to ensure that speech recognition technology is accessible and applicable to diverse language populations.

https://doi.org/10.1051/shsconf/202419703002

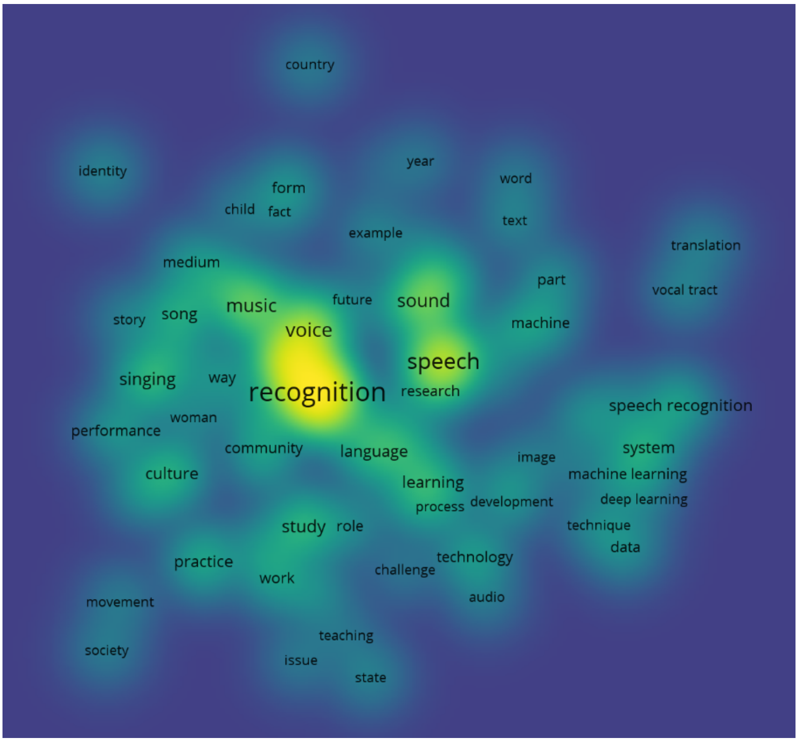

Fig. 2. Density Visualization for machine learning and Indonesian traditional vocal recognition areas.

Machine learning algorithms have become the primary catalyst for advancing vocal recognition in technology. Scientists are always improving and creating new algorithms to address the ever-changing difficulties posed by voice recognition. Researchers have successfully tackled technical challenges such as noise, accents, and multi-modal identification systems, leading to significant breakthroughs. Vocal recognition is becoming more prevalent in education, providing novel resources for language learners and educators. Language comprehension tools that utilize vocal recognition technology offer tailored learning experiences, enhancing the effectiveness and accessibility of education. Vocal recognition systems rely on words and text to translate spoken language into written form. Recent studies have investigated methods to improve the precision of word identification and the comprehension of natural language, thereby closing the divide between spoken and written communication.

The time dimension, represented by the keyword "year," emphasizes the dynamic nature of this topic. Vocal recognition technology progresses significantly each year, offering increasingly advanced applications and solutions. These studies demonstrate a dedication to being at the cutting edge of innovation and ensuring that voice recognition systems stay pertinent and efficient in a constantly evolving world. Singing, an exceptional mode of vocal articulation, has discovered its specialized position within research. Scientists are now studying methods for identifying singing voices, annotating music, and analyzing it, which helps enhance our comprehension of vocal expression within art and culture.

Vocal recognition study also focuses on the cultural and communal aspects, highlighting how technology might influence various sociocultural environments. These investigations aim to guarantee that vocal recognition technology acknowledges and adjusts to the diverse array of human languages, dialects, and cultural customs, thereby promoting inclusiveness and accessibility.

4 Conclusion

The last five years have seen a resurgence in machine learning and voice recognition research, marked by a continuous and collaborative investigation into its various uses and capacities. As technology advances, these studies will influence the future of voice recognition systems, affecting how people communicate, interact, and learn in a digitally interconnected world. The complex interaction among historical, musical, auditory, linguistic, communal, cultural, technological, challenging, pedagogical, anatomical, linguistic, textual, temporal, melodic, and rehearsing factors highlights the comprehensive essence of vocal recognition, exposing its capacity to bring about significant changes in a time where voice plays a central role. Preserving and acknowledging Indonesian traditional vocal music in the digital era is challenging yet essential. Machine learning presents a potential opportunity for automating audio recognition, but it is necessary to modify the technology to suit the distinct attributes of these vocal traditions. Moreover, it is crucial to prioritize cultural awareness and foster engagement with indigenous communities. By employing an appropriate methodology and incorporating advanced machine learning techniques, Indonesia has the potential to effectively safeguard and commemorate its multifarious vocal legacy, guaranteeing the perpetuation of admiration and recognition for this artistic expression among forthcoming generations.

The authors declare that there is no conflict of interest regarding the publication of this article. The authors confirmed that the paper was free of plagiarism.

References

1. C. Cooney, R. Folli, D. Coyle, A bimodal deep learning architecture for EEG-fNIRS decoding of overt and imagined speech. IEEE Trans. Biomed. Eng. **69**, 1983-1994 (2021)
2. O. Balan, A. Moldoveanu, F. Moldoveanu, Navigational audio games: an effective approach toward improving spatial contextual learning for blind people. Int. J. Disabil. Hum. Dev. **14**, 109-118 (2015)
3. R.A. Khalil, E. Jones, M.I. Babar, T. Jan, M.H. Zafar, T. Alhussain, Speech emotion recognition using deep learning techniques: A review. IEEE Access **7**, 117327-117345 (2019)
4. J. Zhang, Z. Yin, P. Chen, S. Nichele, Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review. Inf. Fusion **59**, 103-126 (2020)
5. A. Minks, From children's song to expressive practices: old and new directions in the ethnomusicological study of children. Ethnomusicology **46**, 379-408 (2002)

6. S. Williams, The urbanization of Tembang Sunda, an aristocratic musical genre of West Java, Indonesia, Ph.D. thesis, University of Washington (1990)
7. H.I. Midyanti, R. Tila, E.J. Jaohari, J. Masunah, Design of Soundscape Music on Beluk Vocal in Digitizing Audio Archives, in Fifth International Conference on Arts and Design Education (ICADE 2022) (Atlantis Press, 2023), pp. 498-508
8. I.N. Sedana, K. Foley, The education of a Balinese dalang. Asian Theatre J. **10**, 81-100 (1993)
9. K.Y. Baker, Kecak “Monkey chant” and authenticity in Balinese culture. Found Sounds: UNCG Musicol. J. **2**, (2016)
10. M. Hijleh, Towards a global music history: intercultural convergence, fusion, and transformation in the human musical story (Routledge, 2018)
11. J.C. Kuipers, Language, identity, and marginality in Indonesia: The changing nature of ritual speech on the island of Sumba. Cambridge University Press, 18 (1998)
12. M.J. Rossano, The essential role of ritual in the transmission and reinforcement of social norms. Psychol. Bull. **138**, 529 (2012)
13. P.O.R. Reséndiz, Digital preservation of sound recordings. Investig. Bibliotecológica: Archivonomía, Bibliotecología e Información **30**, 173-195 (2016)
14. P. Conway, Preservation in the digital world (Council on Library and Information Resources, 1996)
15. J.F. Hollifield, The emerging migration state 1. Int. Migr. Rev. 38, 885-912 (2004)
16. D. Huron, Tone and voice: A derivation of the rules of voice-leading from perceptual principles. Music Percept. **19**, 1-64 (2001)
17. S. Suryati, Planning arrangement of medley regional songs on choir for the preservation of local culture. Linguist. Cult. Rev. **5**, 977-991 (2021)
18. T. Baltrušaitis, C. Ahuja, L.P. Morency, Multimodal machine learning: A survey and taxonomy. IEEE Trans. Pattern Anal. Mach. Intell. **41**, 423-443 (2018)
19. K. Lee, J. Nam, Learning a joint embedding space of monophonic and mixed music signals for singing voice. arXiv preprint arXiv:1906.11139 (2019)
20. R.A. Kambau, Z.A. Hasibuan, M.O. Pratama, Classification for multiformat object of cultural heritage using deep learning, in 2018 Third International Conference on Informatics and Computing (ICIC), IEEE (2018), pp. 1-7
21. F.W. Wibowo, Detection of Indonesian Dangdut Music Genre with Foreign Music Genres Through Features Classification Using Deep Learning, in 2021 International Seminar on Machine Learning, Optimization, and Data Science (ISMODE), IEEE (2022), pp. 313-318
22. M. Térmens, Preservación digital (2014)
23. J. Candás-Romero, El papel de los metadatos en la preservación digital. El Prof. Inf. **15**, 126-136 (2006)
24. IASA (International Association of Sound and Audiovisual Archives), Guidelines on the Production and Preservation of Digital Audio Objects. TC-04 (UNESCO, 2006)
25. R. Green, Memoria y preservación digital, in Memorias del Tercer Seminario Internacional. La Preservación de la memoria audiovisual en la sociedad digital, P. Rodríguez, Ed. (Radio Educación, México, 2006)
26. B.E. Asogwa, Digitization of archival collections in Africa for scholarly communication: Issues, strategies, and challenges. Libr. Philos. Pract. **1** (2011)

27. G. Chowdhury, From digital libraries to digital preservation research: the importance of users and context. J. Doc. 66, 207-223 (2010)
28. Y.P. Wang, M.C.C. Chen, Digitization procedures guideline: integrated operation procedures, in Taiwan e-learning and Digital Archives Program (2010)
29. A.N. Lacuata, Digitization of Library Resources in Higher Education Institutions in La Union, Philippines. Preserv. Digit. Technol. Cult. **49**, 139-158 (2020)
30. F. Bressan, A. Rodà, S. Canazza, F. Fontana, R. Bertani, The safeguard of audio collections: a computer science based approach to quality control--the case of the sound archive of the arena di verona. Adv. Multimedia 2013, 7 (2013)
31. M.P. Satija, M. Bagchi, D. Martínez-Ávila, Metadata management and application. Libr. Herald **58**, 84-107 (2020)
32. L. Shklar, A. Sheth, V. Kashyap, K. Shah, InfoHarness: Use of automatically generated metadata for search and retrieval of heterogeneous information, in Advanced Information Systems Engineering: 7th International Conference, CAiSE'95 Jyväskylä, Finland, June 12–16, 1995 Proceedings 7 (Springer Berlin Heidelberg, 1995), pp. 217-230
33. P.O. Rodriguez, Digital preservation of sound recordings. Investig. Bibliotecológica: Archivonomía, Bibliotecología e Información **30**, 173-195 (2016)
34. G. Pessach, The political economy of digital cultural preservation. Digital archives: management, use and access/ur. Milena Dobreva. Facet **39**, 39-72 (2018)
35. T.M. Mitchell, J.G. Carbonell, R.S. Michalski, G. Dejong, A brief overview of explanatory schema acquisition, in Machine Learning: A Guide to Current Research (Springer, 1986), pp. 47-50
36. T.M. Mitchell, Does machine learning really work?. AI Mag. **18**, 11-11 (1997)
37. 7. R. Hecht-Nielsen, Theory of the backpropagation neural network, in Neural networks for perception (Academic Press, 1992), pp. 65-93
38. T. Evgeniou, M. Pontil, Support vector machines: Theory and applications, in Machine Learning and Its Applications: Advanced Lectures (Springer Berlin Heidelberg, 2001), pp. 249-257
39. I. Goodfellow, Y. Bengio, A. Courville, Deep learning (MIT Press, 2016)
40. M.I. Jordan, T.M. Mitchell, Machine learning: Trends, perspectives, and prospects. Science **349**, 255-260 (2015)
41. S.B. Kotsiantis, I. Zaharakis, P. Pintelas, Supervised machine learning: A review of classification techniques. Emerg. Artif. Intell. Appl. Comput. Eng. **160**, 3-24 (2007)
42. S.W. Knox, Machine learning: a concise introduction (John Wiley & Sons, 2018)
43. L.P. Kaelbling, M.L. Littman, A.W. Moore, Reinforcement learning: A survey. J. Artif. Intell. Res. **4**, 237-285 (1996)
44. J. Ma, Machine learning and audio processing: a thesis presented in partial fulfilment of the requirements for the degree of Doctor of Philosophy in Computer Science at Massey University, Albany, Auckland, New Zealand (Massey University, 2019)
45. F. Vesperini, L. Gabrielli, E. Principi, S. Squartini, Polyphonic sound event detection by using capsule neural networks. IEEE J. Sel. Top. Signal Process. **13**, 310-322 (2019)
46. Y. Wang, F. Metze, Connectionist temporal localization for sound event detection with sequential labeling, in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), May 2019, pp. 745-749

47. C.C. Chiu, T.N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, M. Bacchiani. State-of-the-art speech recognition with sequence-to-sequence models, in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), April 2018, pp. 4774-4778
48. S. Petridis, T. Stafylakis, P. Ma, F. Cai, G. Tzimiropoulos, M.S. Pantic, End-to-end audiovisual speech recognition, in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), April 2018, pp. 6548-6552
49. C. Weng, J. Cui, G. Wang, J. Wang, C. Yu, D. Su, D.C. Yu, Improving Attention Based Sequence-to-Sequence Models for End-to-End English Conversational Speech Recognition, in Interspeech, September 2018, pp. 761-765
50. H. Bilen, B. Fernando, E. Gavves, A. Vedaldi, Action recognition with dynamic image networks. IEEE Trans. Pattern Anal. Mach. Intell. **40**, 2799-2813 (2017)
51. H. Rahmani, A. Mian, M. Shah, Learning a deep model for human action recognition from novel viewpoints. IEEE Trans. Pattern Anal. Mach. Intell. **40**, 667-681 (2017)
52. H. Choi, K. Cho, Y. Bengio, Fine-grained attention mechanism for neural machine translation. Neurocomputing **284**, 171-176 (2018)
53. D. He, Y. Xia, T. Qin, L. Wang, N. Yu, T.Y. Liu, W.Y. Ma. Dual learning for machine translation. Adv. Neural Inf. Process. Syst. (NIPS) 820-828 (2016)
54. J. Lee, K. Cho, T. Hofmann, Fully character-level neural machine translation without explicit segmentation. Trans. Assoc. Comput. Linguist. **5**, 365-378 (2017)
55. J. Schmidhuber, Deep learning in neural networks: An overview. Neural Netw. **61**, 85-117 (2015)
56. Y.Y. Song, L. Ying, Decision tree methods: applications for classification and prediction. Shanghai Arch. Psychiatry **27**, 130 (2015)
57. Y.A. Malkov, D.A. Yashunin, Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. IEEE Trans. Pattern Anal. Mach. Intell. **42**, 824-836 (2018)
58. J. Tang, Y. Tian, P. Zhang, X. Liu, Multiview privileged support vector machines. IEEE Trans. Neural Netw. Learn. Syst. **29**, 3463-3477 (2017)
59. B. Du, W. Xiong, J. Wu, L. Zhang, L. Zhang, D. Tao. Stacked convolutional denoising auto-encoders for feature representation. IEEE Trans. Cybern. **47**, 1017-1027 (2016)
60. B. Hutchinson, L. Deng, D. Yu, Tensor deep stacking networks. IEEE Trans. Pattern Anal. Mach. Intell. **35**, 1944-1957 (2012)
61. G.E. Dahl, D. Yu, L. Deng, A. Acero, Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. IEEE Trans. Audio Speech Lang. Process. **20**, 30-42 (2011)
62. S. Ren, K. He, R. Girshick, X. Zhang, J. Sun, Object detection networks on convolutional feature maps. IEEE Trans. Pattern Anal. Mach. Intell. **39**, 1476-1481 (2016)
63. L. Wu, J.Z. Cheng, S. Li, B. Lei, T. Wang, D. Ni, FUIQA: fetal ultrasound image quality assessment with deep convolutional networks. IEEE Trans. Cybern. **47**, 1336-1349 (2017)
64. A. Graves, A.R. Mohamed, G. Hinton, Speech recognition with deep recurrent neural networks. IEEE International Conference on Acoustics, Speech and Signal Processing (2013), pp. 6645-6649
65. M. Sundermeyer, H. Ney, R. Schlüter, From feedforward to recurrent LSTM neural networks for language modeling. IEEE/ACM Trans. Audio Speech Lang. Process. **23**, 517-529 (2015)

66. J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, T. Darrell, Long-term recurrent convolutional networks for visual recognition and description. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2015), pp. 2625-2634
67. G. Heigold, H. Ney, R. Schluter, S. Wiesler, Discriminative training for automatic speech recognition: Modeling, criteria, optimization, implementation, and performance. IEEE Signal Process. Mag. 29, 58-69 (2012)
68. D. Marin, M. Tang, I.B. Ayed, Y. Boykov, Kernel clustering: Density biases and solutions. IEEE Trans. Pattern Anal. Mach. Intell. **41**, 136-147 (2017)
69. L. Huang, H.Y. Chao, C.D. Wang, Multi-view intact space clustering. Pattern Recognit. **86**, 344-353 (2019)
70. I.A. Maraziotis, S. Perantonis, A. Dragomir, D. Thanos, K-Nets: Clustering through nearest neighbors networks. Pattern Recognit. **88**, 470-481 (2019)
71. Y. Yi, J. Wang, W. Zhou, Y. Fang, J. Kong, Y. Lu, Joint graph optimization and projection learning for dimensionality reduction. Pattern Recognit. **92**, 258-273 (2019)
72. C. Örnek, E. Vural, Nonlinear supervised dimensionality reduction via smooth regular embeddings. Pattern Recognit. **87**, 55-66 (2019)
73. M. Harandi, M. Salzmann, R. Hartley, Dimensionality reduction on SPD manifolds: The emergence of geometry-aware methods. IEEE Trans. Pattern Anal. Mach. Intell. **40**, 48-62 (2017)
74. A. Romero, C. Gatta, G. Camps-Valls, Unsupervised deep feature extraction for remote sensing image classification. IEEE Trans. Geosci. Remote Sens. **54**, 1349-1362 (2015)
75. P.M. Sheridan, C. Du, W.D. Lu, Feature extraction using memristor networks. IEEE Trans. Neural Netw. Learn. Syst. **27**, 2327-2336 (2015)
76. Y.A. Ghassabeh, F. Rudzicz, H.A. Moghaddam, Fast incremental LDA feature extraction. Pattern Recognit. **48**, 1999-2012 (2015)
77. A. Hyvarinen, H. Morioka, Unsupervised feature extraction by time-contrastive learning and nonlinear ICA. Adv. Neural Inf. Process. Syst. **29** (2016)
78. C. Doersch, A. Gupta, A.A. Efros, Unsupervised visual representation learning by context prediction. Proceedings of the IEEE International Conference on Computer Vision (2015), pp. 1422-1430
79. L. Tran, X. Yin, X. Liu, Representation learning by rotating your faces. IEEE Trans. Pattern Anal. Mach. Intell. **41**, 3007-3021 (2018)
80. R. Xu, D. Wunsch, Survey of clustering algorithms. IEEE Trans. Neural Netw. **16**, 645-678 (2005)
81. C. Rasmussen, The infinite Gaussian mixture model. Adv. Neural Inf. Process. Syst. **12**, 554-560 (1999)
82. B. Jian, B.C. Vemuri, Robust point set registration using Gaussian mixture models. IEEE Trans. Pattern Anal. Mach. Intell. **33**, 1633-1645 (2010)
83. C.Y. Liou, W.C. Cheng, J.W. Liou, D.R. Liou, Autoencoder for words. Neurocomputing **139**, 84-96 (2014)
84. A. Ghosh, V. Kulharia, V.P. Namboodiri, P.H. Torr, P.K. Dokania, Multi-agent diverse generative adversarial networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018), pp. 8513-8521

85. A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, A.A. Bharath, Generative adversarial networks: An overview. IEEE Signal Process. Mag. **35**, 53-65 (2018)
86. A.K. Jain, M.N. Murty, P.J. Flynn, Data clustering: a review. ACM Comput. Surv. **31**, 264-323 (1999)
87. A.K. Jain, Data clustering: 50 years beyond K-means. Pattern Recognit. Lett. **31**, 651-666 (2010)
88. H.S. Park, C.H. Jun, A simple and fast algorithm for K-medoids clustering. Expert Syst. Appl. **36**, 3336-3341 (2009)
89. Y. Zhao, G. Karypis, U. Fayyad, Hierarchical clustering algorithms for document datasets. Data Min. Knowl. Discov. **10**, 141-168 (2005)
90. H.P. Kriegel, P. Kröger, J. Sander, A. Zimek, Density-based clustering. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. **1**, 231-240 (2011)
91. C. Xu, D. Tao, C. Xu, A survey on multi-view learning. arXiv preprint arXiv:1304.5634 (2013)
92. Y. Li, M. Yang, Z. Zhang, A survey of multi-view representation learning. IEEE Trans. Knowl. Data Eng. 31, 1863-1883 (2018)
93. W. Zhuge, C. Hou, Y. Jiao, J. Yue, H. Tao, D. Yi, Robust auto-weighted multi-view subspace clustering with common subspace representation matrix. PLoS One 12, e0176769 (2017).
94. C. Zhang, Q. Hu, H. Fu, P. Zhu, X. Cao, Latent multi-view subspace clustering. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2017), pp. 4279-4287
95. G. Chao, S. Sun, J. Bi, A survey on multi-view clustering. arXiv preprint arXiv:1712.06246 (2017)
96. Y. Fan, J. Liang, R. He, B.G. Hu, S. Lyu, Robust localized multi-view subspace clustering. arXiv preprint arXiv:1705.07777 (2017)
97. M. Längkvist, L. Karlsson, A. Loutfi, A review of unsupervised feature learning and deep learning for time-series modeling. Pattern Recognit. Lett. **42**, 11-24 (2014)
98. S. Dargan, M. Kumar, M.R. Ayyagari, G. Kumar, A survey of deep learning and its applications: a new paradigm to machine learning. Arch. Comput. Methods Eng. **27**, 1071-1092 (2020)
99. P. Verma, C. Chafe, J. Berger, One-Shot Acoustic Matching Of Audio Signals--Learning to Hear Music In Any Room/Concert Hall. arXiv preprint arXiv:2210.15750 (2022)
100. Y.D. Mistry, G.K. Birajdar, A.M. Khodke, Time-frequency visual representation and texture features for audio applications: a comprehensive review, recent trends, and challenges. Multimed. Tools Appl. 1-35 (2023)
101. A. Ezhilan, R. Dheekksha, S. Shridevi, Audio style conversion using deep learning. Int. J. Appl. Sci. Eng. **18**, 1-8 (2021)
102. J. Chaki, Pattern analysis based acoustic signal processing: a survey of the state-of-art. Int. J. Speech Technol. **24**, 913-955 (2021)
103. L.J.C. Cohen, Using spectral analysis in the flute studio to develop tone quality (Doctoral dissertation, The University of Iowa) (2021)
104. M. Liu, J. Shi, Z. Li, C. Li, J. Zhu, S. Liu, Towards better analysis of deep convolutional neural networks. IEEE Trans. Vis. Comput. Graph. **23**, 91-100 (2016)

105. A. Khan, A. Sohail, U. Zahoora, A.S. Qureshi, A survey of the recent architectures of deep convolutional neural networks. Artif. Intell. Rev. **53**, 5455-5516 (2020)
106. J. Pons, O. Slizovskaia, R. Gong, E. Gómez, X. Serra, Timbre analysis of music audio signals with convolutional neural networks. In 2017 25th European Signal Processing Conference (EUSIPCO) (2017), pp. 2744-2748
107. H. Chaurasiya, Time-frequency representations: spectrogram, cochleogram and correlogram. Procedia Comput. Sci. **167**, 1901-1910 (2020)
108. A. Graves, S. Fernández, J. Schmidhuber, Multi-dimensional recurrent neural networks. In International Conference on Artificial Neural Networks (2007), pp. 549-558
109. E. Cakır, G. Parascandolo, T. Heittola, H. Huttunen, T. Virtanen, Convolutional recurrent neural networks for polyphonic sound event detection. IEEE/ACM Trans. Audio Speech Lang. Process. **25**, 1291-1303 (2017)
110. S.P. Yadav, S. Zaidi, A. Mishra, V. Yadav, Survey on machine learning in speech emotion recognition and vision systems using a recurrent neural network (RNN). Arch. Comput. Methods Eng. **29**, 1753-1770 (2022)
111. C. Gao, J. Yan, S. Zhou, P.K. Varshney, H. Liu, Long short-term memory-based deep recurrent neural networks for target tracking. Inf. Sci. **502**, 279-296 (2019)
112. I.C. Kaadoud, N.P. Rougier, F. Alexandre, Knowledge extraction from the learning of sequences in a long short term memory (LSTM) architecture. Knowl.-Based Syst. **235**, 107657 (2022)
113. E. Tsalera, A. Papadakis, M. Samarakou, Comparison of pre-trained CNNs for audio classification using transfer learning. J. Sensor Actuator Netw. **10**, 72 (2021)
114. S. Shin, J. Kim, Y. Yu, S. Lee, K. Lee, Self-supervised transfer learning from natural images for sound classification. Appl. Sci. **11**, 3043 (2021)
115. A. Triantafyllopoulos, B.W. Schuller, The role of task and acoustic similarity in audio transfer learning: Insights from the speech emotion recognition case. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2021), pp. 7268-7272
116. J. Abeßer, A review of deep learning based methods for acoustic scene classification. Appl. Sci. **10**, 2020 (2020)
117. D. Michelsanti, Z.H. Tan, S.X. Zhang, Y. Xu, M. Yu, D. Yu, J. Jensen, An overview of deep-learning-based audio-visual speech enhancement and separation. IEEE/ACM Trans. Audio Speech Lang. Process. **29**, 1368-1396 (2021)
118. D. de Benito-Gorron, A. Lozano-Diez, D.T. Toledano, J. Gonzalez-Rodriguez, Exploring convolutional, recurrent, and hybrid deep neural networks for speech and music detection in a large audio dataset. EURASIP J. Audio Speech Music Process. **2019**, 1-18 (2019)
119. Z. Zhao, Q. Li, Z. Zhang, N. Cummins, H. Wang, J. Tao, B.W. Schuller, Combining a parallel 2D CNN with a self-attention Dilated Residual Network for CTC-based discrete speech emotion recognition. Neural Netw. 141, 52-60 (2021)