





Review

Multimodal Age and Gender Estimation for Adaptive Human-Robot Interaction: A Systematic Literature Review

Hussain A. Younis ^{1,2}, Nur Intan Raihana Ruhaiyem ^{1,*}, Ameer A. Badr ³, Alia K. Abdul-Hassan ⁴, Ibrahim M. Alfadli ⁵, Weam M. Binjumah ⁶, Eman A. Altuwaijri ⁷ and Maged Nasser ¹

¹ School of Computer science, Universiti Sains Malaysia, Penang 11800, Malaysia; hussain.younis@uobasrah.edu.iq (H.A.Y.)

² College of Education for Women, University of Basrah, Basrah 61004, Iraq

³ Department of Information Technology, Technical College of Management-Baghdad, Middle Technical University, Baghdad 10011, Iraq; ameer.badr@duc.edu.iq

⁴ Department of Computer Science, University of Technology, Baghdad 10011, Iraq

⁵ College of Computer Science and Engineering, Taibah University, Madina 42353, Saudi Arabia; ialfadli@taibahu.edu.sa

⁶ Applied Collage, Taibah University, Madina 42353, Saudi Arabia

⁷ College of Applied Studies and Community Service, King Saud University, Riyadh 145111, Saudi Arabia

* Correspondence: intanraihaana@usm.my

Abstract: Identifying the gender of a person and his age by way of speaking is considered a crucial task in computer vision. It is a very important and active research topic with many areas of application, such as identifying a person, trustworthiness, demographic analysis, safety and health knowledge, visual monitoring, and aging progress. Data matching is to identify the gender of the person and his age. Thus, the study touches on a review of many research papers from 2016 to 2022. At the heart of the topic, many systematic reviews of multimodal pedagogies in Age and Gender Estimation for Adaptive were undertaken. However, no current study of the theme concerns connected to multimodal pedagogies in Age and Gender Estimation for Adaptive Learning has been published. The multimodal pedagogies in four different databases within the keywords indicate the heart of the topic. A qualitative thematic analysis based on 48 articles found during the search revealed four common themes, such as multimodal engagement and speech with the Human-Robot Interaction life world. The study touches on the presentation of many major concepts, namely Age Estimation, Gender Estimation, Speaker Recognition, Speech recognition, Speaker Localization, and Speaker Gender Identification. According to specific criteria, they were presented to all studies. The essay compares these themes to the thematic findings of other review studies on the same topic such as multimodal age, gender estimation, and dataset used. The main objective of this paper is to provide a comprehensive analysis based on the surveyed region. The study provides a platform for professors, researchers, and students alike, and proposes directions for future research.

Keywords: multimodal; age estimation; gender estimation; speech; image; dataset



Citation: Younis, H.A.; Ruhaiyem, N.I.R.; Badr, A.A.; Abdul-Hassan, A.K.; Alfadli, I.M.; Binjumah, W.M.; Altuwaijri, E.A.; Nasser, M. Multimodal Age and Gender Estimation for Adaptive Human-Robot Interaction: A Systematic Literature Review. *Processes* **2023**, *11*, 1488. <https://doi.org/10.3390/pr11051488>

Academic Editors: Adel Ali Ahmed, AbdulRahman Alsewari, Yousef Fazea and Waleed Ali

Received: 18 February 2023

Revised: 6 April 2023

Accepted: 10 April 2023

Published: 15 May 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

With the world's latest tremendous development, technological advancement, and information age, we started accessing this study which is referred to the main terms in the study, i.e., each of Multimodal Age Estimation [1,2]. Typically, it is more challenging to assume the age of a speaker based on their speech [3], Gender Estimation [4], and Human-Robot Interaction [5–7]. The first term, Multimodal, refers to the theory of communication between social auditors that represent communication between audiovisual, visual, and spatial resources. The second term is age and gender estimation. Therefore, many studies have presented Multimodal in meta-learning [8], English language [9], a comprehensive presentation of Vocal sacs [10], deep learning fields of vision, language, and speech [11–15].

One study [16] used three-dimensional image analysis using mandibular third molar apices to estimate age which also includes gender, race, and age image style transfer [17]. The age and gender estimations were based on a large database [18], in which the Gender determination was conducted based on pain [19]. Age and gender estimation were conducted by sphenoid bone pterygoid processes [20] and the Deep Residual Learning Network [21]. In addition, the analysis of race was conducted in [22] the Frequency domain [23]. Infants' neural speech processing [24], brain aging [25,26], and old speakers' ability to understand one another [27]. Emotion recognition [28–30]. The conversion of audio to textured visuals to recognize speech emotions [31,32]. The third term of the Human-Robot Interaction (HRI) is a promising technology in service, social, and industrial perspectives. Therefore, developers of HRI systems should evaluate the effectiveness of the proposed system in terms of its function and whether it will satisfy individual, group, and production requirements (i.e., its quality). Engineering places a lot of emphasis on the concept of quality, which generally indicates how well a system, service, product, or service provider process complies with established requirements and operates within established parameters and conditions. [33–43].

Physical human-robot interaction has emerged as a critical issue in many fields, including human-robot collaboration learning by demonstration and rehabilitation [33,39–45].

Multimodal is a mode of communication that includes Sign Language, Writing, Typing, Body Language, Sign Language, Speech & Vocalizations, Gestures, Facial Expressions, High Tech AAC systems, Light Tech Devices, and Low-Tech boards.

Biometrics is used as an identifier for an individual using physiological data or behavioural attributes. Everyone is known to be unique, therefore, biometric identifiers are permanently associated with the user. As a result, they are more trustworthy than token or knowledge-based authentication techniques, as illustrated in Figure 1. There are three classifications of biometric modalities such as:

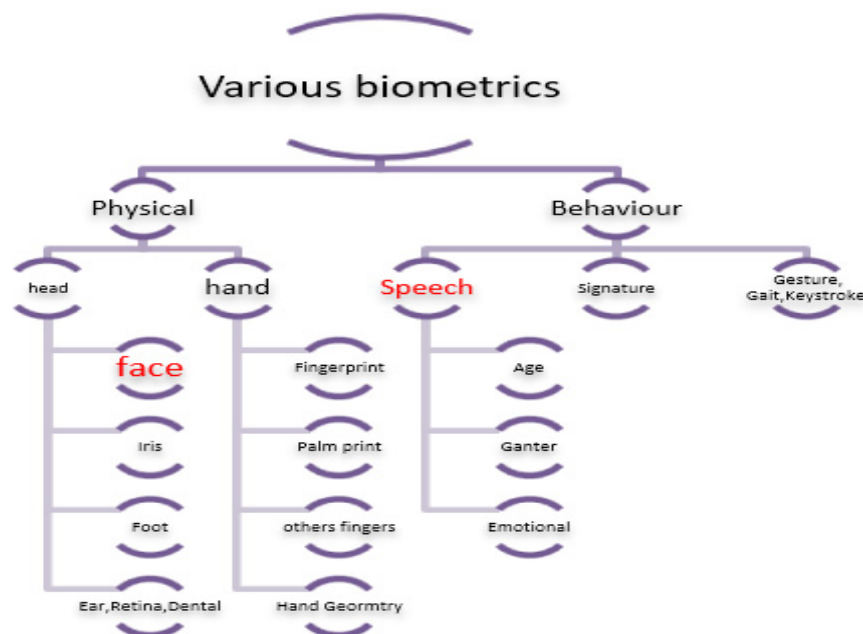


Figure 1. Various Biometrics.

1.1. Physical Biometrics

The unique modality of an individual includes face, fingerprints, iris scans, and hand geometry (see Figure 1). [46].

1.2. Chemical Biometrics

One of the recent areas that involves the estimation of chemical components associated with humans, such as odor and perspiration.

1.3. Behavioral Biometrics

These are commonly temporal in nature and involve analyzing how a user accomplishes a specified task. Examples of modalities include speech, signature, functional mobility, and keyboard dynamics.

This study represented a comprehensive presentation of the study of literature, with 48 discreet scientific articles within a specific period. It is a comprehensive guide and assistant for researchers in the same field that included symptoms of technologies, data, processing methods, analysis, and extraction of features in four important databases, Web of Science database, IEEE Xplore, Science Direct, and Taylor & Francis Group, (25%, 12%, 56%, 6%), respectively.

This paper is outlined as the following, Section 2 Relevant Works represents and compares our A Systematic Literature Review study to the existing literature studies. Section 3 is the Method used in SRL and provides the concept of the study and Research Methodology. Section 4 Results. Section 5, evaluation metrics includes the categories of the studies and discusses a comprehensive and detailed, Section 6, challenges and limitations, Section 7, view of the dataset used in this field; and Section 8 Conclusions of paper review.

2. Relevant Works

This section discusses the existing literature under the scope of multimodal age detection and gender estimation associated with our main research. The main topics included are current systematic reviews, surveys, and reviews. In this work, statistics pertaining to the use of the reviewed studies were also listed (e.g., features, research design, feature selection, deep learning, dataset, speech-dependent, text-dependent, multimodal, language, protocol, and feature direction). The authors also discussed a variety of age and gender estimates and offered some recommendations for potential future research areas, although only in part.

Table 1 represents and compares our A Systematic Literature Review study to the existing literature studies.

Table 1. Comparison between the proposed survey and existing literature.

Survey Paper	Year	Features	Research Design	Feature Selection	Deep Learning	Dataset	Speech Dependent	Text Dependent	Multimodal	Language	Protocol	Feature Direction
[47]	2018	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
[48]	2019	✓	×	×	×	✓	✓	×	✓	×	×	×
[49]	2019	✓	✓	×	×	✓	✓	×	✓	×	×	×
[6]	2020	✓	×	×	×	✓	✓	×	✓	✓	×	×
[50]	2020	✓	×	✓	✓	✓	✓	×	✓	×	×	✓
[51]	2020	✓	✓	×	×	✓	✓	×	✓	✓	×	✓
[52]	2020	✓	✓	×	×	×	×	×	✓	×	×	×
[53]	2020	✓	✓	✓	✓	✓	✓	✓	✓	×	✓	×
[54]	2020	✓	✓	×	×	✓	✓	×	×	×	✓	×
[5]	2021	✓	✓	×	×	✓	✓	×	✓	×	×	×
[55]	2021	✓	✓	×	×	✓	✓	×	✓	×	×	×
[56]	2021	✓	✓	×	✓	✓	✓	×	✓	✓	✓	✓
[57]	2022	✓	✓	×	✓	✓	✓	×	✓	×	✓	×
This Survey	2022	✓	✓	✓	✓	✓	✓	×	✓	✓	✓	✓

Table 1 describes the significance of our work against other existing work in terms of various metrics, such as features, research design, feature selection, deep learning, dataset,

speech-dependent, text-dependent, multimodal, language, protocol, and feature direction. The authors studied and reviewed the body of knowledge on age and gender estimation before establishing these parameters. This work includes a comprehensive discussion of the most recent developments in the field and is compared to 13 prior relevant studies. The scope of the study shows that speech recognition is the extraction of data that represents the identity of the speaking person, to become the person who can, through his voice, reach a set of services, control systems, or databases.

3. Method

This section may be divided by subheadings. It should provide a concise and precise description of the experimental results, their interpretation, as well as the experimental conclusions that can be drawn.

This study designed a group of questions. The study covered the review of inputs, methods of treatment, analysis, and directing of features to the results. We also offered some of the techniques and algorithms. The methods used in the process contribute to each study. The methods were distinguished by the collection of properties and methods of treatment and access, and the presentation of the offer is the most accurate result.

The requirements asserted in the Preferred Reporting Items for Systematic Reviews and Meta-Analyses must be fulfilled (PRISMA) [58–62]. The current study was carried out, meanwhile, PRISMA advises against depending solely on a database. When doing a systematic review, look for the literature because it is unlikely that any database would include all pertinent references. Consequently, much study is practically required, and it is possible to conduct an accurate and detailed survey on several databases to include a variety of articles. This study covers five comprehensive questions for each:

- RQ1: What methods were used and what techniques in data analysis are age estimation and gender knowledge?
- RQ2: Does the data depend on the multi-means of the species? What kind of data is used to estimate age and known gender? Do you rely on multimedia?
- RQ3: What are the challenges? Types, methods of conclusion, challenge it, classify it. Ways to overcome them, how to overcome the challenges?
- RQ4: What essential features of the datasets employed in this investigation are there? Do their traits seem to have an impact on the outcomes?
- RQ5: What are the probable difficulties that exist in the studies in developing an adaptive human-robot interaction's multimodal age and gender estimation?

The most pivotal keywords, which are listed below in italics, describe the scope of this study. *Multimodal, Multimodality, Multimedia, Multi-media, Combined, System can recognize, Age and Gender Estimation, Gender Age and Estimation, estimate, rating, assessment, appreciation, respect, Adaptive, Adapt, Adjust, fit, conform, Human-Robot Interaction, interaction man-robot, Robot-Human Interaction, Data mining, Deep learning, Speech, Voice, and sound.* Specifically, our study is limited to English-language publications. Our study included only selected digital databases such as Scopus, the Web of Science (WoS) database, Taylor & Francis Group, IEEE Xplore, and Science Direct. This research was conducted using our academic search engine accounts to gain access to all open and closed access papers. The first iteration excluded all duplicated articles and selected only articles published prior to the last five and a half years, from 2016 to 2022 (completed June 2022). In this study, the Mendeley-Desktop 1.19.4-win32 was used to collect the papers. The second iteration sorted the articles by titles, abstracts, and articles outside of our domain. The third iteration filtered all articles by reading the full text. Unrelated articles and those that did not meet our research requirements were eliminated.

In this study, we searched the Web of Science (WoS) database, IEEE, Science Direct, and Taylor & Francis Group, using the search string ((*Multimodal OR Multimodality OR Multimedia OR Multimedia OR Combined and System can recognize*) AND (*Age and Gender Estimation OR Gender Age and Estimation*) AND (*estimate OR rating OR assessment OR appreciation OR respect*) AND (*Adaptive OR Adapt OR Adjust*

OR fit OR conform) AND (Human-Robot Interaction OR interaction man-robot OR Robot-Human Interaction) AND (Data mining OR Deep learning)) AND (Speech OR Voice OR sound)). We could not find any Systematic Literature Reviews during our thorough investigation between 2016–2022 that focuses on Age and Gender Estimation for Adaptive Human-Robot Interaction.

3.1. Properties Criteria Search (Eligibility Requirements)

We started to input keywords, and the words of their approach to them in the same sense, to reveal all articles in the same field and within the specified period (see Table 2). Any article does not apply to the conditions below. We included the papers in Figure 2 that matched our criteria based on our analysis. The primary goal is to address the overall scope of the study. The coarse-grained taxonomy has two levels. Categories identified by prior research in the unrestricted literature (only three categories were emphasized in this study). We examined two websites to study the directions cited in the literature. After that, we excluded the articles through three cycles in order to eliminate duplicate papers across databases. When the items failed to fulfil our eligibility requirements for filtering and screening, the following conditions had to be met in order for an article to be disqualified: (1) it must be clear and focus on a single issue; (2) it must be written in English; (3) it must be redundant with other research websites, one of which had to be disqualified; and (4) it must concentrate on the overview. Therefore, we emphasize our efforts on the domains discovered in Figure 2 since they include a wealth of information relevant to the investigation.

Table 2. Original words and alternative words used extensively in the study.

Original Words	Alternative Words
Multimodal	Multimodality Multimedia Multi-media Efficiently manage Combined gdp System can recognize
Age and Gender Estimation	Age and Gender estimate, Age and Gender rating, Age and Gender assessment, Age and Gender appreciation, Age and Gender estimation, Age and Gender respect, Age appreciation and gender appreciation
Adaptive	Adapt Adjust Fit Conform
Human-Robot Interaction Robot-Human Interaction	Interaction man—robot. The interaction between man is a robot. The interaction between a human robot.
Speech	Voice, sound

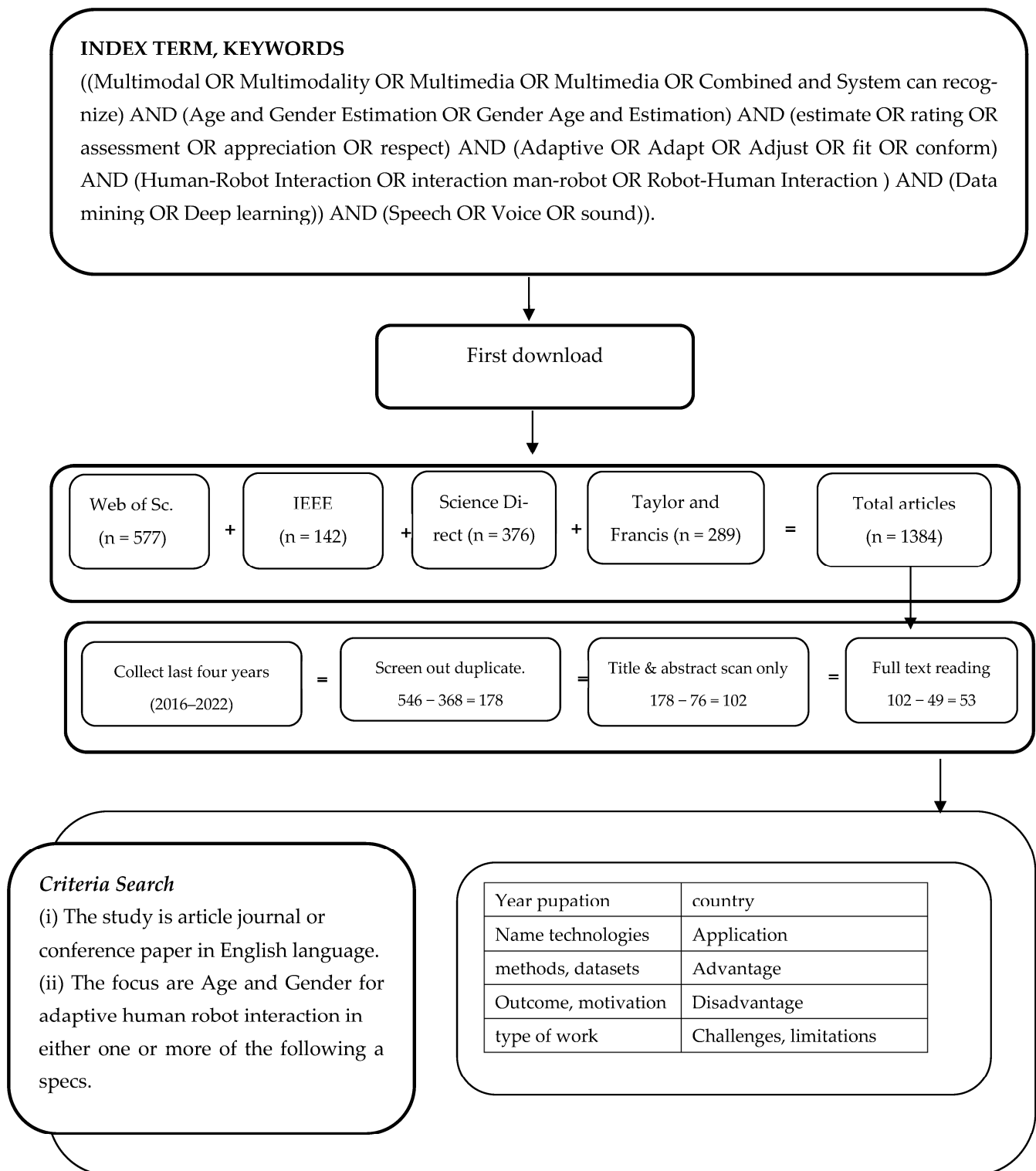


Figure 2. Study flowchart, query terms, and valuation standards.

3.2. Process for Gathering and Taxonomy of Data

The information was gathered from the four scientific websites—WOS, IEEE Xplore, Science Direct, and Taylor & Francis—and secured using the authors' user credentials. The preliminary screening process was then carried out, consisting of the following three steps: First, duplicates are eliminated, then titles and abstracts are examined, and finally, the full texts are reviewed. Last but not least, the Mendeley-Desktop-1.19.4-win32 software, Word,

and Excel were used to establish initial classifications, summaries, and infographics based on their materials after reading, evaluating, and summarizing the articles. All academic articles that followed the taxonomy were divided into two groups. Then, it concentrated on three major categories, including the 48 research papers, models, parameters, type, input, output, average accuracy, and problems, that extensively addressed the issue of multimodal Age and Gender Estimation. We will talk about the other categories in another article.

4. Results

The study recommended that each of the Web of Science database (577), IEEE Xplore (142), Science Direct (376), and Taylor and Francis Group (289), obtain a total of 1384 articles from research sites that were approved in the study between 2016 and 2022. However, two scientific articles from WoS could not be downloaded due to accessibility issues. A total of 546 papers were published in the most recent six years (from 2016 to 2022), and 368 papers appeared in all three databases, generating 178 papers. The articles were filtered using two categories and the research's sequence. Following a careful examination of their titles and abstracts, 67 more papers were eliminated. A total of 53 papers were in the complete set. Five additional research papers were not uploaded after that. As a result, the study now includes 48 academic articles in its entire sample. The study covered two major fields that were divided into two subfields: Multimodal Age and Gender Estimation, and Human-Robot Interaction. The first category consists of 13 papers (27.083%) such as review, survey, and systematic review. Meanwhile, the second category consists of 35 papers (72.91%) that include techniques and algorithms used for age and gender estimation for adaptive Human-Robot Interaction.

4.1. Distribution Outcomes

Figure 3a demonstrates that the four digital databases published many publications. IEEE Xplore published six papers, Web of Science published 12, Science Direct published twenty-seven, and Taylor & Francis published three. Next, we extracted the percentage number of articles from the total number of features in four important databases, Web of Science database, IEEE Xplore, Science Direct, and Taylor & Francis Group (25%, 12%, 56%, 6%), respectively, shown Figure 3b.

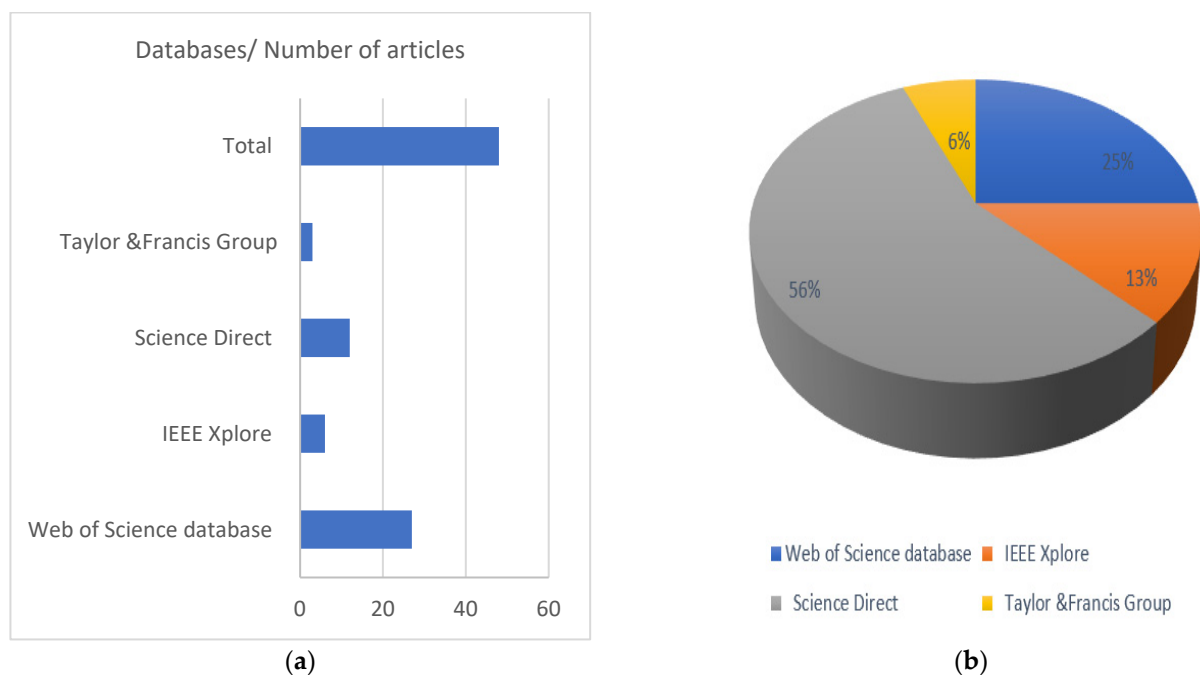


Figure 3. (a) Publishing sites by number (b). The percentage of the number of articles from the total number.

4.2. Distribution by Articles' Publication Years

Figure 4 shows the distribution of the articles from 2016 to 2022 by the years of their publication. Approximately, there have been five published since 2016. In 2017, four articles were published. Since 2018, 23 articles have been published. Since 2019, one article has been published, three have been published since 2020, ten have been published since 2021, and two have been published until May 2022. The lack of articles created in 2019 and 2020 can be attributed to the spread of COVID-19, while the year 2022 included the first four months of it.

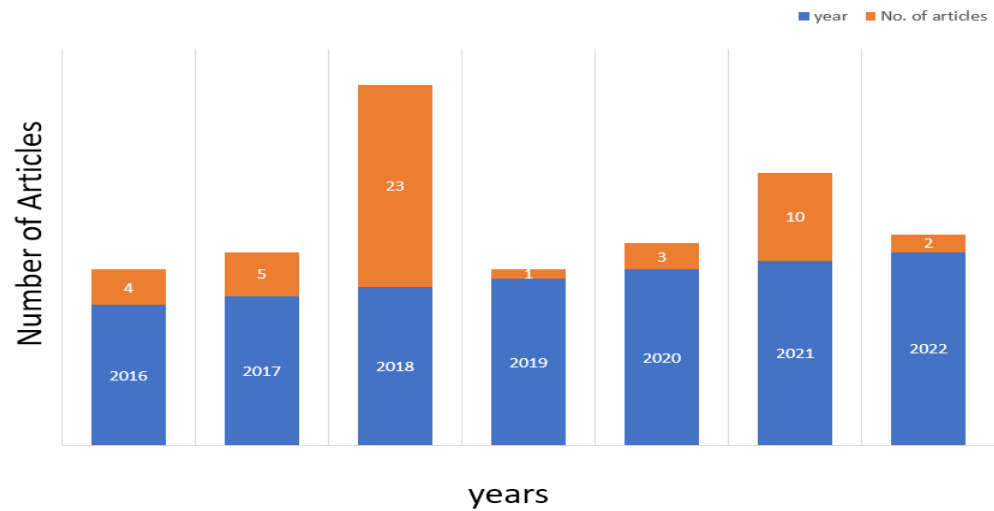


Figure 4. Statistics of the appearance of the number of articles for years.

Figure 5 shows how triage is used in 25 different countries and nationalities. We discovered that certain countries were the focus of literature studies or the circumstances they addressed. The most productive writers come from the United Kingdom (UK), Germany, the United States (USA), Brazil, Australia, Malaysia, Spain, Japan, Iraq, Indonesia, Italy, Turkey, Norway, Iran, Switzerland, Canada, Hungary, India, Francis, Mexico, Chile, Singapore, South Korea, and Switzerland, in that order, in terms of both numbers and percentages (1 each).

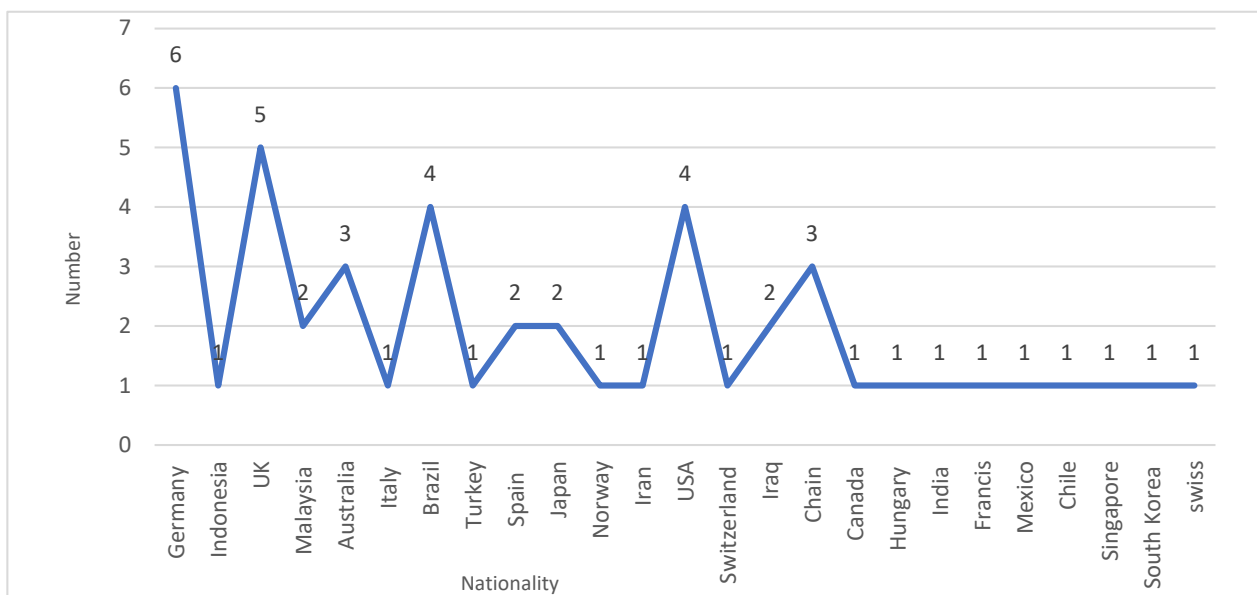


Figure 5. The authors' nationality.

In Figure 6a, the word that appears the biggest in a word cloud reflects the term that occurs the most frequently in the dataset, whereas the word that appears the smallest in a word cloud represents the word that occurs the least frequently in the dataset. While it represents Figure 6b, which was developed to assist instructors in managing the demands of academic language and vocabulary in their text resources.

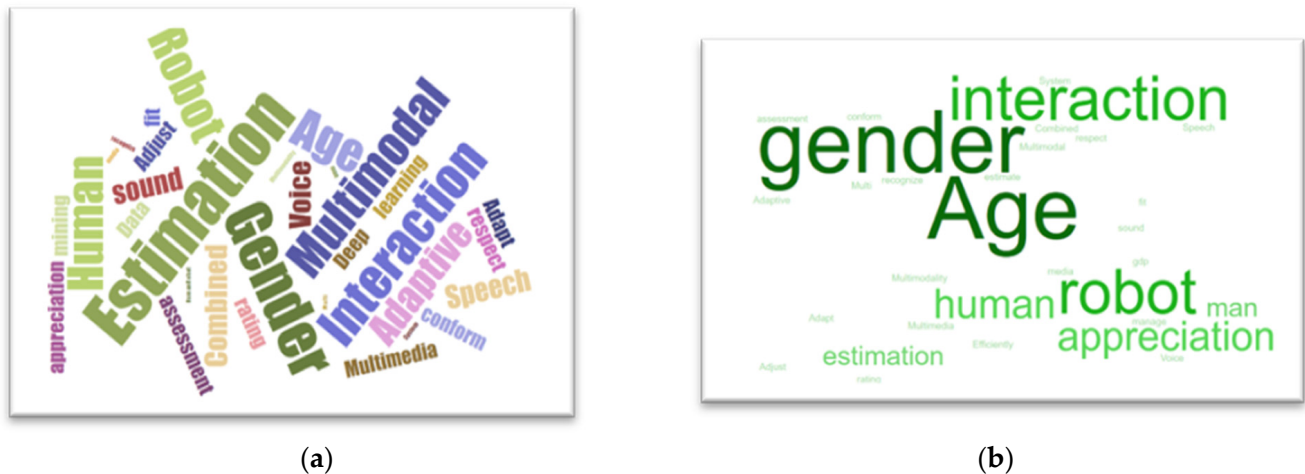


Figure 6. (a) Word frequency (word cloud) in the study in terms of importance, (b) Word Sift.

4.3. Classification

Studies differed from one individual to another by the method of treating inputs, and by what are the approved parameters on which the study was based in terms of input, model, parameters, types, and many other various factors. Some of them used accurate and other learning algorithms. Before entering the topic, it is necessary to include an introduction to deep learning, machine learning, and many methods of age and gender estimation.

4.4. Artificial Intelligence (AI)

Today, artificial intelligence (AI) is known as a fast-expanding discipline that is engaged in various subdomains and adapted to many practical applications. Over decades, AI has been associated with various applications such as self-driving cars, speech and image recognition, machine translation, as well as competing against world chess champions [63].

Some AI systems can learn without being explicitly programmed by extracting patterns from raw data [64]. This capacity is known as Machine Learning (ML). AI-based systems learn through trial and error and develop over time, in contrast to conventional programming that follows step-by-step coding instructions based on logic, if-then rules, and decision trees [65]. A subsidiary of machine learning called deep learning (DL) makes use of artificial neural networks (ANN) with many hidden layers.

A. Machine learning (ML)

Computer algorithms that can learn to perform tasks more effectively based on experience are the subject of machine learning (ML). Essentially, it is related to statistical analysis and pattern recognition. ML has become increasingly mathematical and successful in applications over the last 20 years (Makridakis, Spyr). ML has a wide range of applications, including web searches, commercials, credit scores [66,67], stock market forecasting [68], gene sequencing analysis, behavioral analyses, time forecasting, and in the analysis of large amounts of data. Artificial neural networks have been used to predict time series in recent decades due to ideal properties that allow for nonlinear models to be used [69]. Similarly, the artificial network continues to grow in terms of the development of applications that make it easier to work with when performing simulations with networks.

B. Deep learning (DL)

This is known as an automatic learning technique that uses many layers of nonlinear information processing to extract and transform characteristics with and without supervision, as well as analyze and classify patterns.

An artificial neuron is modeled after a biological structure that serves as a basic information processing unit. An artificial neuron is made up of a linear combination of weights and input values that is processed by activation functions. As shown in Figure 7a, the classical artificial neuron is a binary classifier from a supervised learning algorithm [70].

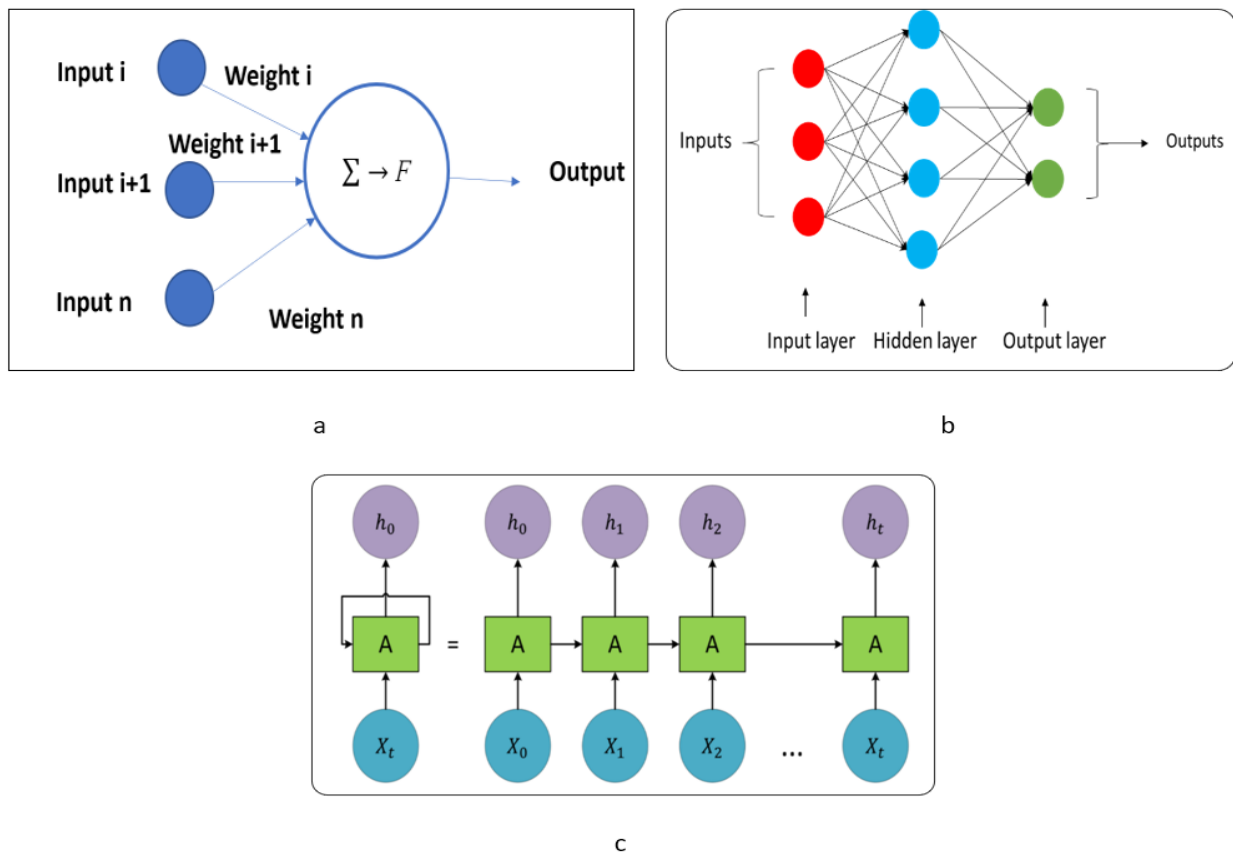


Figure 7. (a) Artificial Neural Network, (b) Neural Network with Three Layers [71,72] (c) Sequential Processing of a Recurrent Neural Network [71,72].

The simplest ANNs generate backward propagation, which allows for better learning patterns to be obtained over time. Therefore, ANNs can match a certain pattern (some pattern of data) over time. f represents the activation function of this unit, which corresponds to the function chosen to convert input x to output value $y = f(x,w)$, and it depends on the specific characteristics of each network.

$$Y_t = f(T_t, S_t, E_t) \tag{1}$$

$$y = f(x, w) \tag{2}$$

where x = input neural network, w = weight.

The basic structure of artificial neurons is a multilayered network. This structure is commonly implemented, and it consists of the first level being the input neuron layers, which receive the values of some of the patterns represented as vectors and act as the input of the network. Then, many hidden layers whose modules respond to specific features may appear on the input patterns; additional hidden layers may be present as well. The final

level is the output, which is presented as the output of the entire network; this network is known as a multi-layered or neural network, (see Equation (3)).

$$f(x) = f^n \left(f^{n-1} \dots \left(f^1(x) \right) \right) \quad (3)$$

where n = number of hidden layers.

The output or forecasting values must be validated by comparing the target to the output values to learn the neural network. The training objective is to validate the forecasting and reduce errors using the loss function. The forecasting is more accurate when the error is lower. Gradient descent is a common improvement approach for improving the loss function.

As presented in Figure 7b, the neural network consists of three primary components (i.e., input layer, hidden layers, and output layer). Firstly, the input layer is used to apply the input data to the network to compute the required results. With one layer, the input layer could have multiple inputs. Secondly, the hidden layers could have one or more layers, depending on the complexity of the network, and they are used to compute the required results as well as train the neural network. Finally, the third layer, which consists of one layer identical to the input layer, is used to generate the computed results. The loss function could be computed to evaluate the network's performance. The forecasting results are compared to the defined target results using the loss function. An example is the mean square error (MSE), which is utilized to compute the difference between the computed and target results. A small difference between the computed and target results refers to the highest accuracy and optimum forecast results, while a large difference refers to the poverty of the technique to produce optimum results that are closely defined to the target. Therefore, the relation between the MSE output and the accuracy is inverse, which means that a small MSE value refers to the optimum forecasting network results. (See Equation (4)).

$$l(y - y') = \frac{1}{n} \sum_{t=1}^n (y - y') \quad (4)$$

where y = actual value (target), y' = forecasting value, and n = number of targets.

Neural networks have an advantage in forecasting due to their adaptive power, but neural networks have limitations in time series because they cannot capture the temporal dependency that could be observed. Thus, RNNs are created.

C. Recurrent neural network (RNN)

RNN was first introduced in the 1980s and is composed of three main layers: an input layer, one or more hidden layers, and an output layer. RNNs have a series structure as repeating units, with the idea of storing and managing information from previous processing steps in these units. In time series, RNN can learn the sequence and solve the dependency problem. Figure 7c shows a simple RNN with an input unit, an output unit, and a recurring hidden unit spread throughout the network, where x_t is the input at t time and h_t is the output at once.

During the training process, RNN utilizes the reaction propagation algorithm that is commonly used in the gradient calculation and modification of the matrices that will be modified after adjusting the feedback process. (See Equations (5) and (6)).

$$h^t = f \left(Ux^t, Wh^{(t-1)} \right) \quad (5)$$

$$o^t = \text{softmax}(Vh^t) \quad (6)$$

For this reason, this algorithm is also commonly called Back-Propagation Through Time (BPTT) [73].

D. Long short-term memory (LSTM)

LSTM is a revolution in RNN. Hochreiter and Schmidhuber (1997) proposed the process to address the discomforts of RNN when additional cells are added, known as the ability to learn long-term dependencies and recall of information for extended periods of time [72,74].

In Table 3, the classification represents studies according to the factors that were mentioned earlier by accreditation, models, and variables to display the main elements of output, accuracy, and challenges within the field of study.

Face-based age estimation algorithms are frequently used in biometric applications and other industries such as forensics and healthcare. To validate or estimate an individual's age for security purposes, facial characteristics can be used to restrict access to physical or logical resources for that person. It will be indicated, clarified, and displayed in Figure 8 in our study.

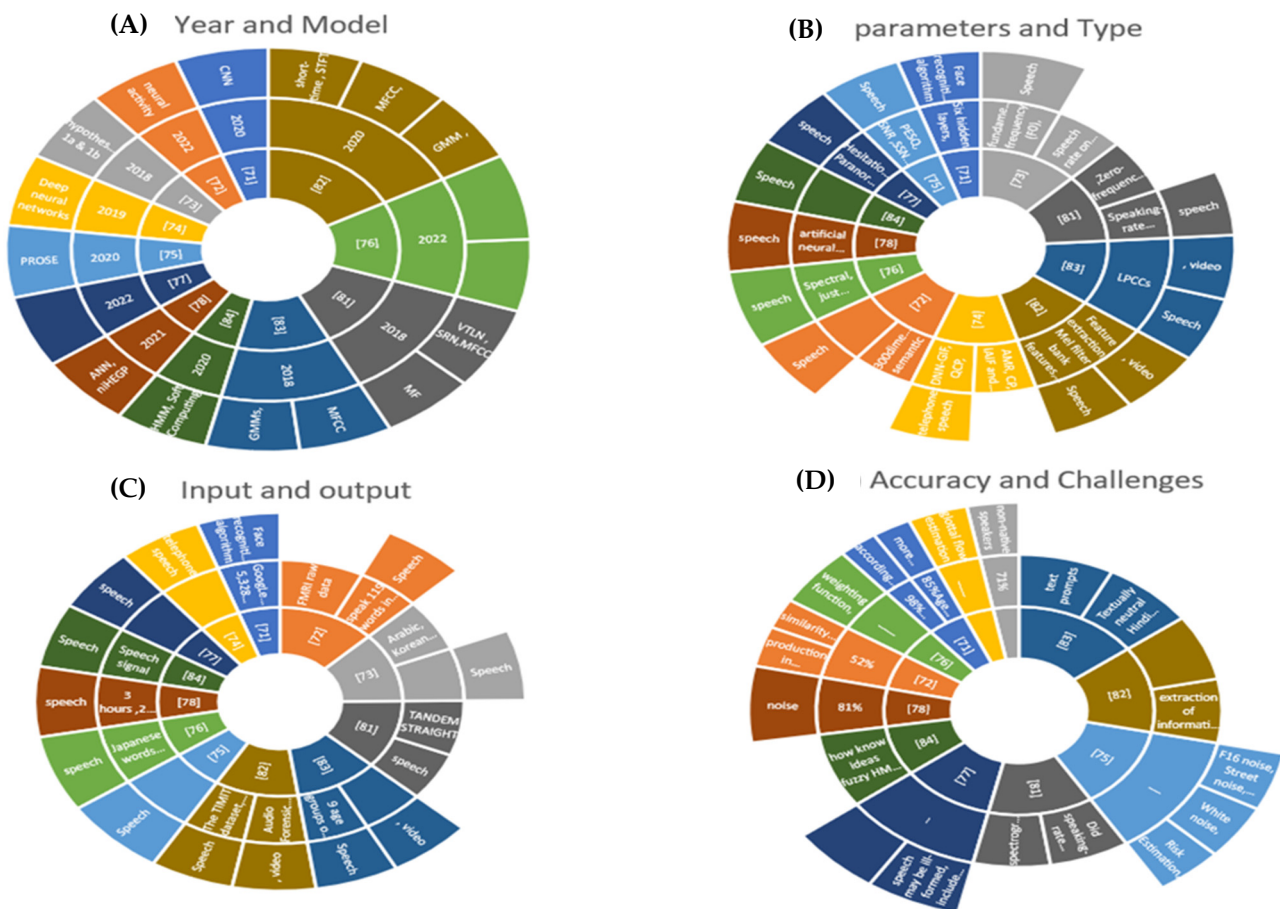


Figure 8. The results of the target category. (A) Reference, Year and Model, (B) References, Parameters and Type, (C) References, Input, and output, and (D) References, Accuracy and Challenges.

Table 3. Classification based on input, model, parameters, types, and other factors.

Ref.	Year	Model	Parameters	Type	Input	Out Put	Average Accuracy	Challenges
[75]	2020	CNN	Six hidden layers,	Face recognition algorithm	GoogLeNet's, 5328 images	Estimation	85%Age est., 98% gender esti. 98%	more difficult to classify the more detailed classes according to age.
[76]	2022	neural activity	Five categories, 300 dimensional semantic	Speech	speak 119 words in the MRI, FMRI raw data	Estimation	52%	similarity & language perception and language production in semantic representations
[77]	2018	hypotheses 1a & 1b	fundamental frequency (F0), Speech rate on age estimates.	Speech	50 participants, Arabic, Korean and Mandarin	Estimation	71%	non-native speakers
[78]	2019	Deep neural networks	DNN-GIF, QCP, AMR, CP, IAIF and CCD.	telephone speech	windowing, interpolate and overlap add	Estimation	—	glottal flow estimation
[79]	2020	PROSE	PESQ, SNR, SSNR, SDR, STOI, MSE, WE, IS, COSH, IS-II	Speech	clean speech NOIZEUS and NOISEX-92 database	Estimation	—	Risk Estimation, F16 noise, White noise, Street noise, Train noise.
[80]	2022	stabilized wavelet transform (SWT), Auditory mode, Cross-correlation, Transform SWM,	Spectral, just noticeable difference (JND, point of subjective equality (PSE),	speech	Japanese words (FW03), TANDEM-STRAIGHT dataset	Estimation	—	weighting function,
[81]	2022	DNN, LSTM	Hesitations, Paranormal, filler words word fragments, word repetition	speech	text augmentation approaches by adult data, Web data, text generated by RNN, 33 children, LENA recording	Estimation, Speech Recognition	—	degree of spontaneity, include incomplete sentences. speech may be ill-formed, vocal effort
[82]	2021	ANN, niHEGP	artificial neural networks	speech	3 h, 2023 newspaper style sentences, 300 sentences [83,84]	Estimation	81%	noise
[85]	2018	MF VTLN, SRN, MFCC	Speaking-rate normalization, Zero-frequency filtering	speech	TANDEM STRAIGHT	Speech Recognition		spectrogram, Did speaking-rate normalization (SRN) highly effective?

Table 3. Cont.

Ref.	Year	Model	Parameters	Type	Input	Out Put	Average Accuracy	Challenges
[86]	2020	GMM, MFCC, short-time, STFT	Feature extraction, Mel filter bank features, UBM, RMSE, MAE	Speech, video	The TIMIT dataset, 630 speakers Audio Forensic Dataset (AFDS)	Estimation		extraction of information, speaker characteristics
[87]	2018	GMMs, MFCC	LPCCs	Speech, video	9 age groups	Classification		Textually neutral Hindi words have been used to construct text prompts
[88]	2020	HMM, Soft Computing	(i) Rules inference engine (ii) Fuzzifier (iii) Defuzzifier	Speech	Speech signal	Speech Recognition, Classification		how know ideas fuzzy HMM strategy, how know speech signal handling territory
[89]	2020	support vector machines (SVMs), time-domain, frequency-domain parameters	glottal parameters: time-domain, frequency-domain parameters, PCA	Speech	dysarthric speech database, 765 isolated word utterances (B1, B2 and B3) 255 words, 155 common words	Classification	72.01% Test data, 73.53% Validation data	classification, intelligibility estimation tasks compared,

In addition, in 2022, a study introducing random auditory stimulation was presented for the final time. Individual perception (RaS-DeeP) used Deep Expectation (DEX CNN) to achieve an EER of 3.3% for age estimation using Deep-learning-based age estimation [90].

On one hand, this aspect will be reviewed and explained by a display of models, dataset entered, versions, and sizes arrived in the results (see Table 3). On the other hand, a study used k-nearest neighbor (KNN) by gait energy images as to age estimation [91] and gait analysis [92].

5. Evaluation Metrics

They are standard criteria that are measured by the actual performance of the work. It is the level that must be the performance and arises usually after observation, monitoring, experiments, research, and testing. These criteria are determined in advance, so, there are two types of accuracy evaluation metrics, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE) are used. In this study, the standards were tested on the news classification of the speaker (female/male) and will be described in detail [1,93–95].

A. Mean Absolute Error (MAE)

Error, also known as MAE, is a metric that measures the number of errors in a forecasting set without taking the direction of the error into account (error greater or less than the true value). This is shown in Figure 9 (See Equation (7)).

$$MAE = \frac{1}{n} \sum_{t=1}^n |y_t - y'_t| \tag{7}$$

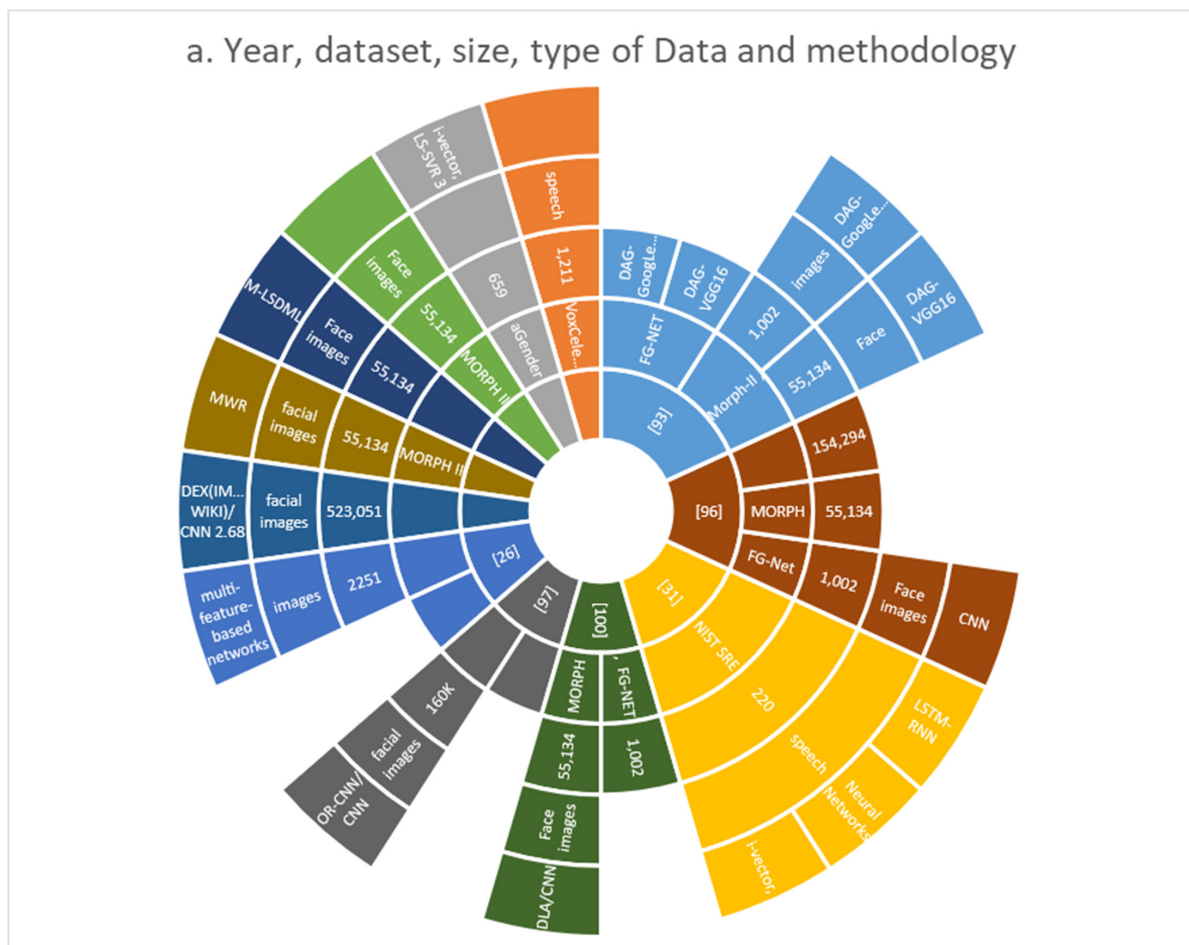


Figure 9. Cont.

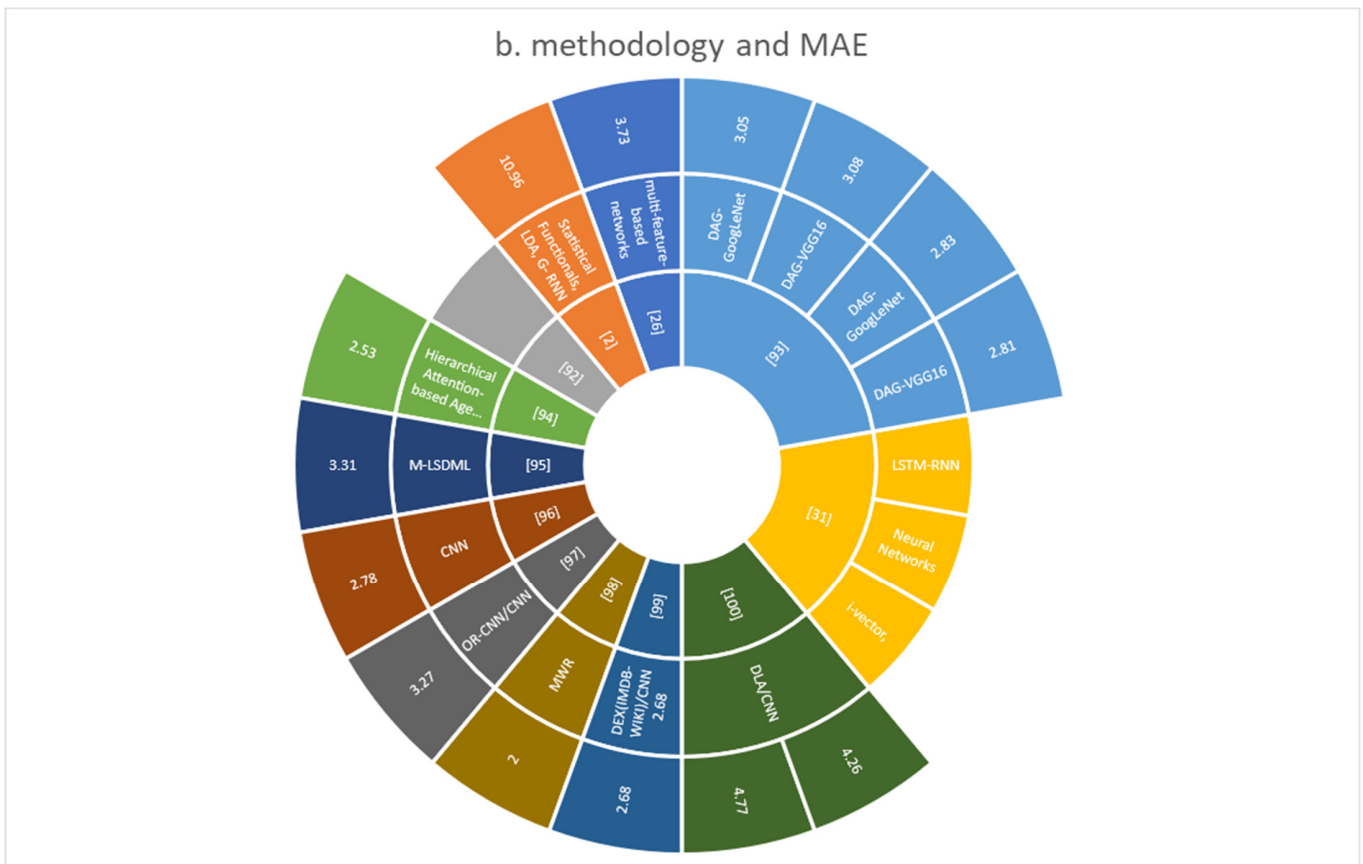


Figure 9. (a) Year, dataset, size, type of data, and methodology, (b) methodology and MAE.

B. Root Mean Square Error (RMSE)

RMSE is a commonly utilized method for calculating the error of a model in forecasting quantitative data (See Equation (8)). It is formally defined as follows:

$$RMSE = \sqrt{\sum_{t=1}^n \frac{(y'_t - y_t)^2}{n}} \tag{8}$$

RMSE or MAE is one of the most used metrics for measuring the mean error between predicted and real values. The RMSE is more sensitive to outliers in the data because it calculates the mean of the squared errors.

6. Challenges and Limitations

Multimodal is a mode of communication that includes Language, Writing, Typing, Body Language, Sign Language, Speech and Vocalizations, Gestures, and Facial Expressions; High Tech AAC systems, Light Tech Devices, and Low-Tech boards.

In Figure 10, there are many challenges in the field of the study of the exact specialization represented by many difficulties which are in the role of finding solutions to them. It can be shortened by capturing the five as shown in Figure 8, which are devices, spatial challenges, time challenges, methods used, and other factors. The first devices contain type of devices, modernity number, and the number of speakers per second and scan image. The second spatial challenges contain the place, nature, delivery, and work environment. The third challenge contains the age of the speaker, their health, and psychology. The fourth method used contain RNN, CNN, DT, and another (see Table 4, Methodology column). Finally, other factors include the used dataset, variables, sample size, and techniques.

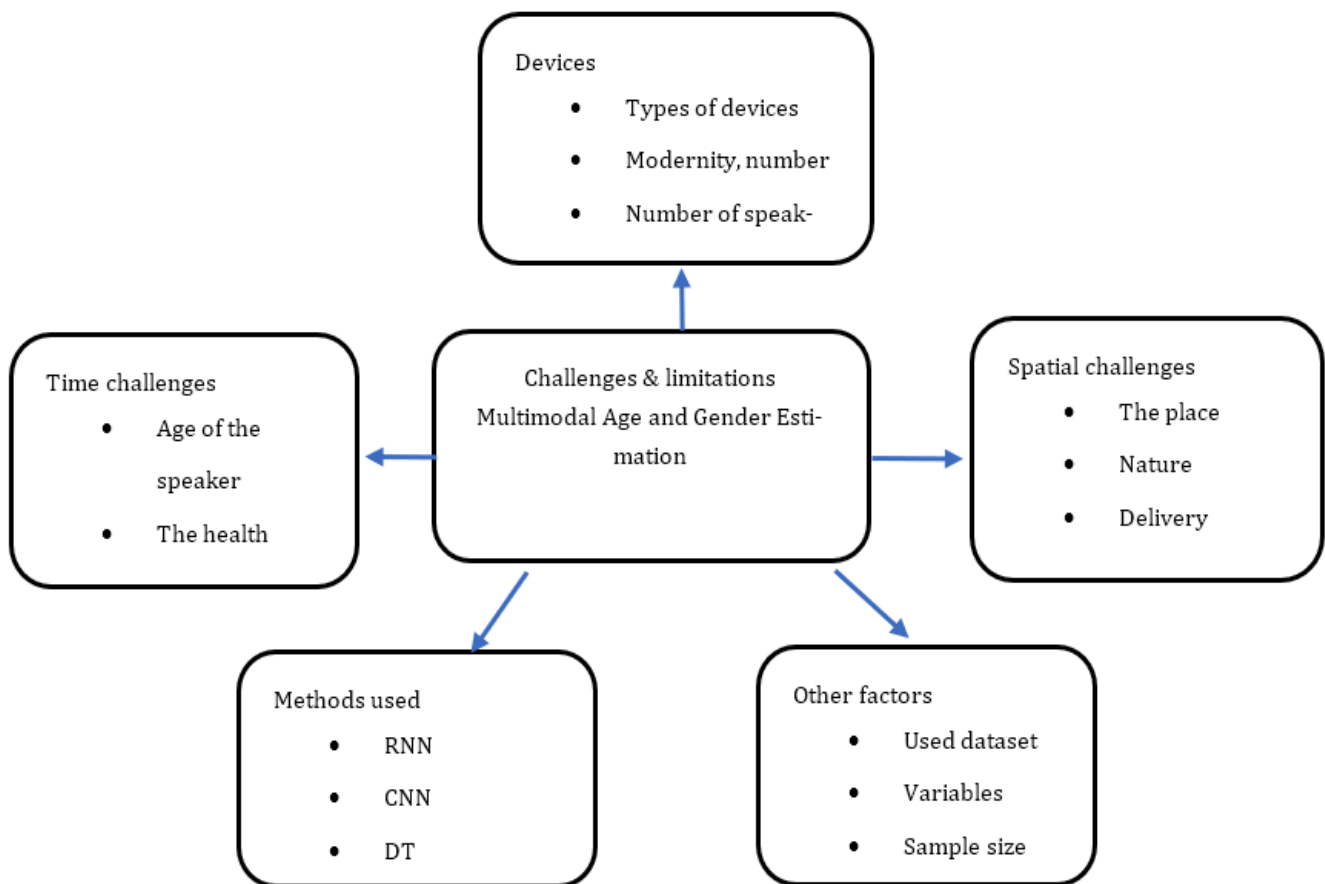


Figure 10. Challenges and Limitations of Age and Gender Estimation Studies.

Table 4. The summary of the related articles, their methodology, used dataset and their details, methodology and MAS (Female, Male and Mixed).

Ref.	Year	Dataset	Size	Type of Data	Methodology	MAE (Years)		
						Female	Male	Mixed
[26]	2022	ADNI, ADNI, DLBS, 1000FCP, IXI, NITRC, OASIS, PPMI, SALD	2251	images	multi-feature-based networks			3.73
[2]	2021	VoxCeleb1	1211	speech	Statistical Functional, LDA, G- RNN	9.25	10.33	10.96
[96]	2016	aGender	659	Telephone Speech	i-vector, LS-SVR 3	9.77	10.63	
[31]	2018	NIST SRE	220	speech	i-vector, Neural Networks LSTM RNN	9.85 9.56 7.44 6.97	10.82 10.69 8.29 7.79	
[97]	2019	Morph-II, FG-NET	55,134 1002	Face images	DAG-VGG16 DAG-GoogLeNet			2.81 2.83
[98]	2021	MORPH II	55,134	Face images	Hierarchical Attention-based Age Estimation			2.53
[99]	2018	MORPH (Album2)	55,134	Face images	M-LSDML			3.31
[100]	2017	FG-Net MORPH Year-labeled.	1002 55,134 154,294	Face images	CNN			2.78
[101]	2016	Asian Face Age Dataset (AFAD)	160 K	facial images	OR-CNN/CNN			3.27
[102]	2022	MORPH II	55,134	facial images	MWR			2.00
[103]	2016	IMDB-WIKI	523,051	facial images	DEX(IMDB-WIKI)/CNN 2.68			2.68
[104]	2015	MORPH, FG-NET	55,134 1002	Face images	DLA CNN			4.77 4.26
[105]	2018	Morph-II	55,134	facial images	CNN + ELM			3.44
[106]	2017	FGNET, MORPH, FERET, PAL	1002, 2000 2366, 576	facial images	CNN			5.39, 3.98 3.00, 3.43
[107]	2016	IMDB-Wiki	500K	Face images	CNN			2.99

7. Dataset

In this section, the data were reviewed. We were able to divide the data into two versions, A and B; the first contains the TIMIT Dataset, the HKUST Dataset, the SRE08/10 Dataset, the TCDSA Dataset, the JNAS Dataset, the AgeVoxCeleb dataset, and the AgeVox-Celeb2 dataset. As for the second group, it contained the utch corpus dataset, the UF-VAD dataset, the aGender dataset, and the NIST SRE dataset. Each of the arguments was reviewed separately, the form of the components of each category, its parts, numbers, types, sizes, and studies that used this category data groups. Mono-media and multi-means.

All data contain different forms and types, including speaking, sound, and others including videos and pictures. All of these are aimed at reaching the assessment of Age and Gender Estimation. It represents real data.

A. Group A Dataset

1. TIMIT DATASET

The purpose of the TIMIT read speech corpus is to develop speech data for acoustic-phonetic research as well as for the development and evaluation of automatic speech recognition systems. Broadband recordings of 630 speakers reading ten phonetically dense phrases in eight significant American English dialects are included in TIMIT. Each utterance in the TIMIT corpus has a 16-bit, 16-kHz speech waveform file, and time-aligned orthographic, phonetic, and word transcriptions. Texas Instruments recorded the lecture, MIT transcribed it, and the National Institute of Standards and Technology inspected and prepared it for CD-ROM production (NIST) [1,2,4,108–110].

2. HKUST DATASET

The Hong Kong University of Science and Technology (HKUST) created HKUST Mandarin Telephone Speech, which comprises roughly 149 h of Mandarin conversational telephone speech (CTS). Although Standard Mandarin is not a natural dialect in the majority of China, it is considered the official language of instruction. Mandarin speakers can have regional accents or not. All calls were audited and classified as standard or accented without further distinction, and subjects' birthplaces were categorized into Mandarin-dominant and non-Mandarin-dominant regions. This dataset was used in several cities across China. Most of the participants had never met before. For this purpose, topics similar to Fisher English were developed to engage in a smooth conversation. All calls were initiated by an automated operator dialing two participants at the scheduled time of contact. Participants were asked age, gender, native language/dialect, education, employment, phone type education, and other demographic questions (refer to Table 5). It denotes the type and number [111–113].

Table 5. Type And Number.

Grope	Number of Calls	Number of Hours	Females	Males
Training	873	144.7	797	948
Development	24	3.9	24	24
Total	867	148.6	821	972

3. SRE08/10 DATASET

There are two types of SRE Datasets, namely the SRE08 Dataset and the SRE10 Dataset. From 2008 to 2010, revisions and modifications were made. The 2008 NIST Speaker Recognition Evaluation Test Set was published in cooperation with the Linguistic Data Consortium (LDC) and NIST (National Institute of Standards and Technology). It consists of 942 h of multilingual telephone conversation, English interview speech, transcripts, and other materials that were utilized as test data in the 2008 NIST Speaker Recognition Evaluation (SRE). The NIST SRE dataset is a sequence of NIST evaluations that are essential in determining the focus of research and the apex of technological development. They are tailored

toward academics with a general interest in text-independent speaker detection. The interview is conducted in English the entire time. About 368 h of the dataset were accounted for by telephone speech, and the remaining 574 h were accounted for by microphone speech [31,114–116].

4. TCDSA DATASET

The Trinity College Dublin Speaker Ageing (TCDSA) Database was created with the aim of learning more about the influence of age-related voice changes on speaker verification. A major part of the collection comprises speech recordings from 26 individuals (15 males and 11 females) whose range of age starts from 25 to 58 years. A collection of 120 developing speakers is also offered, with a mix of ages, genders, and accents. [117,118].

5. JNAS DATASET

The Acoustical Society of Japan's Project the Speech Database Committees. The following are the contents of newspaper article sentences: 155 text sets (each with roughly 100 sentences), for a total of 16176 sentences. The 503 phonetically balanced phrases in ATR are as follows; there are a total of 503 sentences in ten text sets (each with roughly 50 sentences). There are 306 speakers in all (153 males and 153 females). The environment for recording is a head-set microphone, and a desk-top microphone is included. WAV is a digital audio format (16 kHz, 16-bit, Mono) and is formatted for audio files [119–121].

6. AGEVOXCELEB DATASET

The VoxCeleb dataset is an audio-visual collection of short snippets of human speech collected from YouTube video interviews. It has over 7000 clips. Speakers VoxCeleb features almost 2000 h of speech from speakers of all races, accents, professions, and ages. VoxCeleb is a combination of voice and visual apparatus. Each section lasts at least three seconds. There is a total of 1,000,000 utterances. All speaking face tracks, as well as laughing, with background chatter, laughter, different lighting conditions, overlapping speech, and pose variation. Gender was distributed as follows: 61% male and 39% female. The countries of the speakers include the following, the United States, the United Kingdom Germany, India, France, and Unknown, 29%, 10%, 6%, 6%, 6%, 7%, respectively [3,122–125].

7. AGEVOXCELEB2 DATASET

This dataset consists of 5994 speakers, 145569 videos, and 1092009 utterances. It is a giant database of more than a million utterances from 6112 celebrities collected from YouTube videos for VoxCeleb2. The IDs in the VoxCeleb2 development set do not overlap with those in the VoxCeleb1 or SITW datasets [124–126].

B. GROUP B DATASET

1. DUTCH CORPUS DATASET

A huge dual set includes 425 speakers from the N-best training corpus's Flemish section which includes news, interviews, live interviews, read commentary, and reports from Belgium. The corpus contains all of them. Figure 11 displays the age histograms for both male and female speakers [127,128].

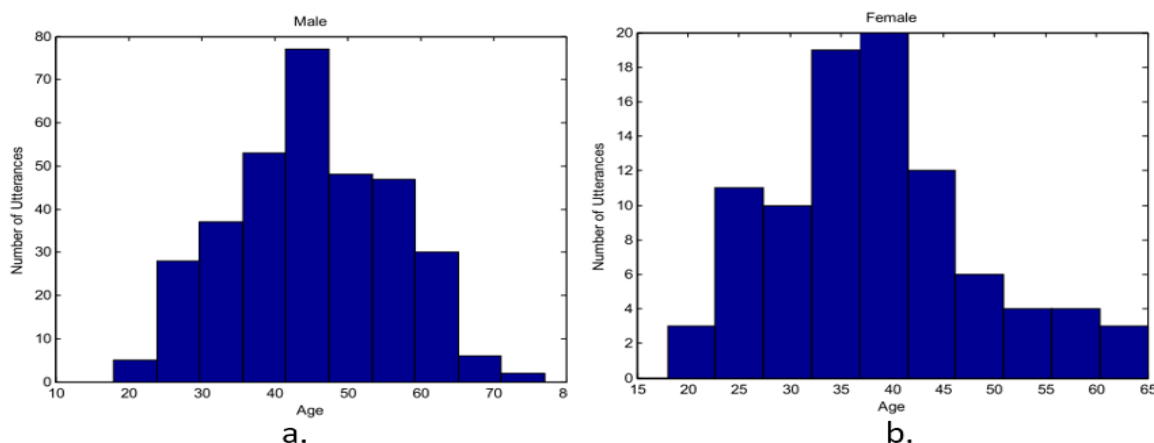


Figure 11. Speakers' histogram (a) male (b) female.

2. UF-VAD DATASET

The University of Florida Vocal Aging Database (UF-VAD), a corpus of American English that was taped between 2003 and 2007, provided the information for the tests. The collection includes 1350 read spoken utterances from 150 different speakers that are based on well-known sources like the Rainbow Passage, Grandfather Passage, and SPIN phrases. A corpus with a duration of 5 h consists of a participant who reads 2 min of the same material using the same microphone and recording settings. The contributing speakers are similarly evenly distributed among the three main age groups and genders. In this instance, there are 25 male and female speakers in the young (18–29), middle-aged (40–55), and elderly (over 55) age brackets (62–92). There are 50 delegates for each age group and 75 delegates for each gender. The average age for each age group is 21, 48, and 79, respectively [129,130].

3. AGENDER DATASET

Felix Burkhardt et al. presented the aGender corpus in A Database of Age and Gender Annotated Telephone Speech, which consists of audio recordings of free speech and predetermined utterances made by people of various ages and genders. As shown in Table 5, each utterance is categorized into one of three gender categories—male, female, or child—as well as one of four age categories—children (C), youth (Y), middle (M), or senior [96,131–134] (Refer to Table 6). The dataset is summarized as 659 speakers (291 males and 368 females).

Table 6. Agender dataset (age and gender).

Class No.	Age	Gender	Age Group
1	7–14	Male + Female	Children
2	15–24	Female	Young
3	15–24	Male	Young
4	25–54	Female	Male
5	25–54	Male	Male
6	55–80	Female	Seniors
7	55–80	Male	Seniors

4. NIST SRE DATASET

The most enormous and comprehensive dataset for telephone speaker recognition currently available is the NIST SRE CTS Superset, which is used to compile earlier SRE datasets (SRE1996-2012). Overall, there are 605,760 segments with 6867 voices (2885 male and 3992 female). As some speakers appear in multiple source corpora, the total number of speakers in the table exceeds 6867. Each speaker has at least three sessions or calls, with each section containing somewhere between 10 and 60 s of speaking. The CTS Superset contains more than 50 languages, even though English is spoken in the vast majority of segments (both native and accented English) [31,135].

8. Conclusions

This work suggested a comprehensive framework literature review where the methods and techniques of several studies were clarified during the study period and show the types of data used in this field. It also identified some studies that used automatic learning and interaction between humans and robots, and the study showed the study environment and displayed MAE. In a systematic study of Age and Gender Estimation, the approach was presented to determine the main challenges and restrictions in the implementation of the double-stage research. We studied accurately many studies to show the benefits, challenges, and recommendations using accurate data, Age and Gender Estimation, and we identified several gaps.

This research provides a broader base for modern literature directions by data analysis, determining their sizes and knowing their type, up to the results of the disposal in research papers in Age and Gender Estimation Studies to obtain the deepest visions in the

investigation area. The goal is to encourage academics and practitioners to use multimedia analysis that includes speech and images and rely on them to reach a lifetime and select sex for a precise image. Moreover, it determines the inputs on which the review operations relied upon for the most accurate results and avoid the amount of error.

Author Contributions: Conceptualization, N.I.R.R. and H.A.Y.; methodology, N.I.R.R. and H.A.Y.; software, N.I.R.R., M.N. and H.A.Y.; validation, I.M.A., W.M.B. and E.A.A.; formal analysis, H.A.Y.; investigation, H.A.Y. and M.N.; resources, H.A.Y. and A.A.B.; data curation, H.A.Y.; writing—original draft preparation, A.A.B. and A.K.A.-H.; writing—review and editing, I.M.A., W.M.B. and E.A.A.; visualization, N.I.R.R. and M.N.; supervision, N.I.R.R., A.A.B. and M.N.; project administration A.A.B. and A.K.A.-H.; funding acquisition N.I.R.R., I.M.A. and W.M.B. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Fundamental Research Grant Scheme FRGS/1/2021/ICT04/USM/02/1/Ministry of Higher Education.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: We would like to express our gratitude to both University of Basrah, Universiti Sains Malaysia (USM), University of Tabuk, University of Technology in Baghdad, Middle Technical University, Taibah University and, King Saud University for all the support and facilities that enable the completion of this research.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

Definition

Percep-Tual Evaluation of Speech Quality
 Segmental Snr
 Source-To-Distortion Ratio
 Short-Time Objective Intelligibility
 Perceptual Risk Optimization For Speech Enhancement
 Mean-Square Error
 Weighted Euclidean Distortion
 Itakura–Saito Distortion
 Hyperbolic Cosine Distortion Measure
 Itakura-Saito Distortion Between Dct Power Spectra
 Stabilised Wavelet Transform
 Just Noticeable Difference
 Point Of Subjective Equality
 Long Short-Term Memory
 Stabilized Wavelet-Melling Transform
 Stabilized Wavelets Transform
 Deep-Neural Network
 Asian Face Age Dataset
 random auditory stimulation and Deep-learning based age estimation, though individual perception
 used Deep Expectation—Convolutional neural network
 Systematic Literature Review
 Preferred Reporting Items for Systematic Reviews and Meta-Analyses
 Artificial Intelligence
 Machine Learning
 Deep Learning
 Recurrent Neural Networks
 Artificial Neural Network

Abbreviations

PESQ
 SSNR
 SDR
 STOI
 PROSE
 MSE
 WE
 IS
 COSH
 IS-II
 SWT
 JND
 PSE
 LSTM
 SWMT
 SWT
 DNN
 AFAD
 RaS-DeeP
 DEX-CNN
 SLR
 PRISMA
 AI
 ML
 DL
 RNN
 ANN

Audio Forensic Dataset	AFDS
Mel filter cepstral coefficients	MFCC
Mean Absolute Error	MAE
short-time Fourier transform	STFT
Gaussian mixture models	GMMs
Hidden Markov Model	HMM
support vector machines	SVMs
principal component analysis	PCA
ADNI: the Alzheimer’s Disease Neuroimaging Initiative,	ADNI
Dallas Lifespan Brain Study	DLBS
Functional Connectomes Project	1000FCP
Information eXtraction from Images	IXI
Neuro Imaging Tools & Resources Collaborator	NITRC
Open Access Series of Imaging Studies	OASIS
Parkinson’s Progression Markers Initiative,	PPMI
label-sensitive deep metric learning	LSDML
Moving Window Regression	MWR
Southwest University Adult Lifespan Dataset	SALD
Mean Absolute Error	MAE
Root Mean Square Error	RMSE

References

1. Badr, A.A.; Abdul-Hassan, A.K. Estimating Age in Short Utterances Based on Multi-Class Classification Approach. *Comput. Mater. Contin.* **2021**, *68*, 1713–1729. [\[CrossRef\]](#)
2. Badr, A.A.; Abdul-Hassan, A.K. Age Estimation in Short Speech Utterances Based on Bidirectional Gated-Recurrent Neural Networks. *Eng. Technol. J.* **2021**, *39*, 129–140. [\[CrossRef\]](#)
3. Minematsu, N.; Sekiguchi, M.; Hirose, K. Automatic estimation of one’s age with his/her speech based upon acoustic modeling techniques of speakers. In Proceedings of the 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing, Orlando, FL, USA, 13–17 May 2002; Volume 1, pp. I-137–I-140.
4. Badr, A.A.; Karim, A. Speaker gender identification in matched and mismatched conditions based on stacking ensemble method. *J. Eng. Sci. Technol.* **2022**, *17*, 1119–1134.
5. Younis, H.; Mohamed, A.; Jamaludin, R.; Wahab, N. Survey of Robotics in Education, Taxonomy, Applications, and Platforms during COVID-9. *Comput. Mater. Contin.* **2021**, *67*, 687–707. [\[CrossRef\]](#)
6. Ayounis, H.; Jamaludin, R.; Wahab, M.; Mohamed, A. The review of NAO robotics in Educational 2014–2020 in COVID-19 Virus (Pandemic Era): Technologies, type of application, advantage, disadvantage and motivation. *IOP Conf. Ser. Mater. Sci. Eng.* **2020**, *928*, 032014. [\[CrossRef\]](#)
7. Younis, H.A.; Mohamed, A.; Ab Wahab, M.N.; Jamaludin, R.; Salisu, S. A new speech recognition model in a human-robot interaction scenario using NAO robot: Proposal and preliminary model. In Proceedings of the 2021 International Conference on Communication & Information Technology (ICICT), Basrah, Iraq, 5–6 June 2021; pp. 215–220. [\[CrossRef\]](#)
8. Ma, Y.; Zhao, S.; Wang, W.; Li, Y.; King, I. Multimodality in meta-learning: A comprehensive survey. *Knowl.-Based Syst.* **2022**, *250*, 108976. [\[CrossRef\]](#)
9. Lim, F.V.; Toh, W.; Nguyen, T.T.H. Multimodality in the English language classroom: A systematic review of literature. *Linguist. Educ.* **2022**, *69*, 101048. [\[CrossRef\]](#)
10. Li, H.; Schrode, K.M.; Bee, M.A. Vocal sacs do not function in multimodal mate attraction under nocturnal illumination in Cope’s grey treefrog. *Anim. Behav.* **2022**, *189*, 127–146. [\[CrossRef\]](#)
11. Shrestha, A.; Mahmood, A. Review of Deep Learning Algorithms and Architectures. *IEEE Access* **2019**, *7*, 53040–53065. [\[CrossRef\]](#)
12. Song, Z.; Yang, X.; Xu, Z.; King, I. Graph-Based Semi-Supervised Learning: A Comprehensive Review. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**, 1–21. [\[CrossRef\]](#)
13. Young, T.; Hazarika, D.; Poria, S.; Cambria, E. Recent Trends in Deep Learning Based Natural Language Processing. *IEEE Comput. Intell. Mag.* **2018**, *13*, 55–75. [\[CrossRef\]](#)
14. Guo, Y.; Liu, Y.; Oerlemans, A.; Lao, S.; Wu, S.; Lew, M.S. Deep learning for visual understanding: A review. *Neurocomputing* **2016**, *187*, 27–48. [\[CrossRef\]](#)
15. Asif, M.K.; Nambiar, P.; Ibrahim, N.; Al-Amery, S.M.; Khan, I.M. Three-dimensional image analysis of developing mandibular third molars apices for age estimation: A study using CBCT data enhanced with Mimics & 3-Matics software. *Leg. Med.* **2019**, *39*, 9–14. [\[CrossRef\]](#)
16. Kim, Y.H.; Nam, S.H.; Hong, S.B.; Park, K.R. GRA-GAN: Generative adversarial network for image style transfer of Gender, Race, and age. *Expert Syst. Appl.* **2022**, *198*, 116792. [\[CrossRef\]](#)
17. Guo, G.; Mu, G. A framework for joint estimation of age, gender and ethnicity on a large database. *Image Vis. Comput.* **2014**, *32*, 761–770. [\[CrossRef\]](#)

18. Zhang, L.; Losin, E.A.R.; Ashar, Y.K.; Koban, L.; Wager, T.D. Gender Biases in Estimation of Others' Pain. *J. Pain* **2021**, *22*, 1048–1059. [[CrossRef](#)]
19. de Sousa, A.L.A.; da Silva, B.A.K.; Lopes, S.L.P.D.C.; Mendes, J.D.P.; Pinto, P.H.V.; Pinto, A.S.B. Estimation of gender and age through the angulation formed by the pterygoid processes of the sphenoid bone. *Forensic Imaging* **2022**, *28*, 200489. [[CrossRef](#)]
20. Lee, S.H.; Hosseini, S.; Kwon, H.J.; Moon, J.; Koo, H.I.; Cho, N.I. Age and gender estimation using deep residual learning network. In Proceedings of the 2018 International Workshop on Advanced Image Technology (IWAIT), Chiang Mai, Thailand, 7–9 January 2018; pp. 1–3. [[CrossRef](#)]
21. Puc, A.; Struc, V.; Grm, K. Analysis of Race and Gender Bias in Deep Age Estimation Models. In Proceedings of the 2020 28th European Signal Processing Conference (EUSIPCO), Amsterdam, The Netherlands, 18–21 January 2021; pp. 830–834. [[CrossRef](#)]
22. Lee, S.S.; Kim, H.G.; Kim, K.; Ro, Y.M. Adversarial Spatial Frequency Domain Critic Learning for Age and Gender Classification. In Proceedings of the 2018 25th IEEE International Conference on Image Processing (ICIP), Athens, Greece, 7–10 October 2018; pp. 2032–2036. [[CrossRef](#)]
23. Zhao, T.C.; Kuhl, P.K. Development of infants' neural speech processing and its relation to later language skills: A MEG study. *Neuroimage* **2022**, *256*, 119242. [[CrossRef](#)]
24. Tremblay, P.; Brisson, V.; Deschamps, I. Brain aging and speech perception: Effects of background noise and talker variability. *Neuroimage* **2020**, *227*, 117675. [[CrossRef](#)]
25. Liu, X.; Beheshti, I.; Zheng, W.; Li, Y.; Li, S.; Zhao, Z.; Yao, Z.; Hu, B. Brain age estimation using multi-feature-based networks. *Comput. Biol. Med.* **2022**, *143*, 105285. [[CrossRef](#)]
26. Zeng, J.; Peng, J.; Zhao, Y. Comparison of speech intelligibility of elderly aged 60–69 years and young adults in the noisy and reverberant environment. *Appl. Acoust.* **2019**, *159*, 107096. [[CrossRef](#)]
27. Arya, R.; Singh, J.; Kumar, A. A survey of multidisciplinary domains contributing to affective computing. *Comput. Sci. Rev.* **2021**, *40*, 100399. [[CrossRef](#)]
28. Maithri, M.; Raghavendra, U.; Gudigar, A.; Samanth, J.; Barua, P.D.; Murugappan, M.; Chakole, Y.; Acharya, U.R. Automated emotion recognition: Current trends and future perspectives. *Comput. Methods Programs Biomed.* **2022**, *215*, 106646. [[CrossRef](#)]
29. Egger, M.; Ley, M.; Hanke, S. Emotion Recognition from Physiological Signal Analysis: A Review. *Electron. Notes Theor. Comput. Sci.* **2019**, *343*, 35–55. [[CrossRef](#)]
30. Zazo, R.; Nidadavolu, P.S.; Chen, N.; Gonzalez-Rodriguez, J.; Dehak, N. Age Estimation in Short Speech Utterances Based on LSTM Recurrent Neural Networks. *IEEE Access* **2018**, *6*, 22524–22530. [[CrossRef](#)]
31. Bakhshi, A.; Harimi, A.; Chalup, S. CyTex: Transforming speech to textured images for speech emotion recognition. *Speech Commun.* **2022**, *139*, 62–75. [[CrossRef](#)]
32. Gustavsson, P.; Syberfeldt, A.; Brewster, R.; Wang, L. Human-robot Collaboration Demonstrator Combining Speech Recognition and Haptic Control. *Procedia CIRP* **2017**, *63*, 396–401. [[CrossRef](#)]
33. Dimeas, F.; Aspragathos, N. Online Stability in Human-Robot Cooperation with Admittance Control. *IEEE Trans. Haptics* **2016**, *9*, 267–278. [[CrossRef](#)]
34. Song, C.S.; Kim, Y.-K. The role of the human-robot interaction in consumers' acceptance of humanoid retail service robots. *J. Bus. Res.* **2022**, *146*, 489–503. [[CrossRef](#)]
35. Cui, Y.; Song, X.; Hu, Q.; Li, Y.; Sharma, P.; Khapre, S. Human-robot interaction in higher education for predicting student engagement. *Comput. Electr. Eng.* **2022**, *99*, 107827. [[CrossRef](#)]
36. Zhang, Q.; Fang, L.; Zhang, Q.; Xiong, C. Simultaneous estimation of joint angle and interaction force towards sEMG-driven human-robot interaction during constrained tasks. *Neurocomputing* **2022**, *484*, 38–45. [[CrossRef](#)]
37. Kim, H.; So, K.K.F.; Wirtz, J. Service robots: Applying social exchange theory to better understand human–robot interactions. *Tour. Manag.* **2022**, *92*, 104537. [[CrossRef](#)]
38. Coronado, E.; Kiyokawa, T.; Ricardez, G.A.G.; Ramirez-Alpizar, I.G.; Venture, G.; Yamanobe, N. Evaluating quality in human-robot interaction: A systematic search and classification of performance and human-centered factors, measures and metrics towards an industry 5.0. *J. Manuf. Syst.* **2022**, *63*, 392–410. [[CrossRef](#)]
39. Paliga, M.; Pollak, A. Development and validation of the fluency in human-robot interaction scale. A two-wave study on three perspectives of fluency. *Int. J. Hum.-Comput. Stud.* **2021**, *155*, 102698. [[CrossRef](#)]
40. Lee, K.H.; Baek, S.G.; Lee, H.J.; Lee, S.H.; Koo, J.C. Real-time adaptive impedance compensator using simultaneous perturbation stochastic approximation for enhanced physical human–robot interaction transparency. *Robot. Auton. Syst.* **2022**, *147*, 103916. [[CrossRef](#)]
41. Secil, S.; Ozkan, M. Minimum distance calculation using skeletal tracking for safe human-robot interaction. *Robot. Comput. Manuf.* **2022**, *73*, 102253. [[CrossRef](#)]
42. Chen, J.; Ro, P.I. Human Intention-Oriented Variable Admittance Control with Power Envelope Regulation in Physical Human-Robot Interaction. *Mechatronics* **2022**, *84*, 102802. [[CrossRef](#)]
43. Liu, H.; Fang, T.; Zhou, T.; Wang, Y.; Wang, L. Deep Learning-based Multimodal Control Interface for Human-Robot Collaboration. *Procedia CIRP* **2018**, *72*, 3–8. [[CrossRef](#)]
44. Grasse, L.; Boutros, S.J.; Tata, M.S. Speech Interaction to Control a Hands-Free Delivery Robot for High-Risk Health Care Scenarios. *Front. Robot. AI* **2021**, *8*, 612750. [[CrossRef](#)]

45. Dargan, S.; Kumar, M. A comprehensive survey on the biometric recognition systems based on physiological and behavioral modalities. *Expert Syst. Appl.* **2020**, *143*, 113114. [CrossRef]
46. Cummins, N.; Baird, A.; Schuller, B.W. Speech analysis for health: Current state-of-the-art and the increasing impact of deep learning. *Methods* **2018**, *151*, 41–54. [CrossRef] [PubMed]
47. Imani, M.; Montazer, G.A. A survey of emotion recognition methods with emphasis on E-Learning environments. *J. Netw. Comput. Appl.* **2019**, *147*, 102423. [CrossRef]
48. Tapus, A.; Bandera, A.; Vazquez-Martin, R.; Calderita, L.V. Perceiving the person and their interactions with the others for social robotics—A review. *Pattern Recognit. Lett.* **2019**, *118*, 3–13. [CrossRef]
49. Badr, A.; Abdul-Hassan, A. A Review on Voice-based Interface for Human-Robot Interaction. *Iraqi J. Electr. Electron. Eng.* **2020**, *16*, 1–12. [CrossRef]
50. Akçay, M.B.; Oğuz, K. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Commun.* **2020**, *116*, 56–76. [CrossRef]
51. Berg, J.; Lu, S. Review of Interfaces for Industrial Human-Robot Interaction. *Curr. Robot. Rep.* **2020**, *1*, 27–34. [CrossRef]
52. Shoumy, N.J.; Ang, L.-M.; Seng, K.P.; Rahaman, D.; Zia, T. Multimodal big data affective analytics: A comprehensive survey using text, audio, visual and physiological signals. *J. Netw. Comput. Appl.* **2019**, *149*, 102447. [CrossRef]
53. Grossi, G.; Lanzarotti, R.; Napoletano, P.; Noceti, N.; Odone, F. Positive technology for elderly well-being: A review. *Pattern Recognit. Lett.* **2020**, *137*, 61–70. [CrossRef]
54. Abdu, S.A.; Yousef, A.H.; Salem, A. Multimodal Video Sentiment Analysis Using Deep Learning Approaches, a Survey. *Inf. Fusion* **2021**, *76*, 204–226. [CrossRef]
55. Fahad, S.; Ranjan, A.; Yadav, J.; Deepak, A. A survey of speech emotion recognition in natural environment. *Digit. Signal Process.* **2021**, *110*, 102951. [CrossRef]
56. Bjørk, M.B.; Kvaal, S.I. CT and MR imaging used in age estimation: A systematic review. *J. Forensic Odonto-Stomatol.* **2018**, *36*, 14–25.
57. Kofod-petersen, A. How to do a structured literature review in computer science. *Researchgate* **2014**, *1*, 1–7.
58. Veras, L.G.D.O.; Medeiros, F.L.L.; Guimaraes, L.N.F. Systematic Literature Review of Sampling Process in Rapidly-Exploring Random Trees. *IEEE Access* **2019**, *7*, 50933–50953. [CrossRef]
59. Keele, S. Guidelines for Performing Systematic Literature Reviews in Software Engineering. Tech. Report, Ver. 2.3 EBSE Tech. Report. EBSE. 2007. Available online: https://www.elsevier.com/_data/promis_misc/525444systematicreviewsguide.pdf (accessed on 12 February 2023).
60. Götz, S. Supporting systematic literature reviews in computer science: The systematic literature review toolkit. In Proceedings of the 21st ACM/IEEE International Conference on Model Driven Engineering Languages and Systems: Companion Proceedings, Proceedings of the MODELS '18: ACM/IEEE 21th International Conference on Model Driven Engineering Languages and Systems, Copenhagen Denmark, 14–19 October 2018; Association for Computing Machinery: New York, NY, USA; pp. 22–26. [CrossRef]
61. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. The PRISMA 2020 Statement: An Updated Guideline for Reporting Systematic Reviews. *BMJ* **2021**, *372*, 105906. [CrossRef]
62. Makridakis, S. The forthcoming Artificial Intelligence (AI) revolution: Its impact on society and firms. *Futures* **2017**, *90*, 46–60. [CrossRef]
63. Lele, A.; Lele, A. Artificial intelligence (AI). Disruptive technologies for the militaries and security. In *Disruptive Technologies for the Militaries and Security*; Springer: Berlin/Heidelberg, Germany, 2019; Volume 132, pp. 139–154.
64. Makridakis, S.; Spiliotis, E.; Assimakopoulos, V. The M4 Competition: Results, findings, conclusion and way forward. *Int. J. Forecast.* **2018**, *34*, 802–808. [CrossRef]
65. Makridakis, S.; Spiliotis, E.; Assimakopoulos, V. The Accuracy of Machine Learning (ML) Forecasting Methods versus Statistical Ones: Extending the Results of the M3-Competition. In *Working Paper, University of Nicosia*; Institute for the Future: Palo Alto, CA, USA, 2017.
66. Hayder, I.M.; Al Ali, G.A.N.; Younis, H.A. Predicting reaction based on customer's transaction using machine learning approaches. *Int. J. Electr. Comput. Eng.* **2023**, *13*, 1086–1096.
67. Wang, J.; Wang, J. Forecasting stochastic neural network based on financial empirical mode decomposition. *Neural Netw.* **2017**, *90*, 8–20. [CrossRef]
68. Kock, A.B.; Teräsvirta, T. Forecasting Macroeconomic Variables Using Neural Network Models and Three Automated Model Selection Techniques. *Econ. Rev.* **2015**, *35*, 1753–1779. [CrossRef]
69. McMahan, H.B.; Ramage, D.; Com, B.G. Federated Learning of Deep Networks using Model Averaging. *arXiv* **2012**, arXiv:1602.05629.
70. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016.
71. Le, X.H.; Ho, H.V.; Lee, G.; Jung, S. Application of Long Short-Term Memory (LSTM) Neural Network for Flood Forecasting. *Water* **2019**, *11*, 1387. [CrossRef]

72. Hayder, I.M.; Al-Amiedy, T.A.; Ghaban, W.; Saeed, F.; Nasser, M.; Al-Ali, G.A.; Younis, H.A. An In-telligent Early Flood Forecasting and Prediction Leveraging Machine and Deep Learning Algorithms with Ad-vanced Alert System. *Processes* **2023**, *11*, 481. [[CrossRef](#)]
73. Zhao, J.; Huang, F.; Lv, J.; Duan, Y.; Qin, Z.; Li, G.; Tian, G. Do RNN and LSTM have long memory? In Proceedings of the 37th International Conference on Machine Learning, ICML, Vienna, Austria, 12–18 July 2020; pp. 11302–11312.
74. Lim, S. Estimation of gender and age using CNN-based face recognition algorithm. *Int. J. Adv. Smart Converg.* **2020**, *9*, 203–211. [[CrossRef](#)]
75. Lin, Y.; Hsieh, P.-J. Neural decoding of speech with semantic-based classification. *Cortex* **2022**, *154*, 231–240. [[CrossRef](#)] [[PubMed](#)]
76. Jiao, D.; Watson, V.; Wong, S.G.-J.; Gnevshva, K.; Nixon, J.S. Age estimation in foreign-accented speech by non-native speakers of English. *Speech Commun.* **2018**, *106*, 118–126. [[CrossRef](#)]
77. Narendra, N.; Airaksinen, M.; Story, B.; Alku, P. Estimation of the glottal source from coded telephone speech using deep neural networks. *Speech Commun.* **2019**, *106*, 95–104. [[CrossRef](#)]
78. Sadasivan, J.; Seelamantula, C.S.; Muraka, N.R. Speech Enhancement Using a Risk Estimation Approach. *Speech Commun.* **2020**, *116*, 12–29. [[CrossRef](#)]
79. Matsui, T.; Irino, T.; Uemura, R.; Yamamoto, K.; Kawahara, H.; Patterson, R.D. Modelling speaker-size discrimination with voiced and unvoiced speech sounds based on the effect of spectral lift. *Speech Commun.* **2022**, *136*, 23–41. [[CrossRef](#)]
80. Lileikyte, R.; Irvin, D.; Hansen, J.H. Assessing child communication engagement and statistical speech patterns for American English via speech recognition in naturalistic active learning spaces. *Speech Commun.* **2022**, *140*, 98–108. [[CrossRef](#)]
81. Tang, Y. Glimpse-based estimation of speech intelligibility from speech-in-noise using artificial neural networks. *Comput. Speech Lang.* **2021**, *69*, 101220. [[CrossRef](#)]
82. Cooke, M. A glimpsing model of speech perception in noise. *J. Acoust. Soc. Am.* **2006**, *119*, 1562–1573. [[CrossRef](#)] [[PubMed](#)]
83. Cooke, M.; Mayo, C.; Valentini-Botinhao, C.; Stylianou, Y.; Sauert, B.; Tang, Y. Evaluating the intelligibility benefit of speech modifications in known noise conditions. *Speech Commun.* **2013**, *55*, 572–585. [[CrossRef](#)]
84. Shahnawazuddin, S.; Adiga, N.; Kathania, H.K.; Pradhan, G.; Sinha, R. Studying the role of pitch-adaptive spectral estimation and speaking-rate normalization in automatic speech recognition. *Digit. Signal Process.* **2018**, *79*, 142–151. [[CrossRef](#)]
85. Kalluri, S.B.; Vijayasanen, D.; Ganapathy, S. Automatic speaker profiling from short duration speech data. *Speech Commun.* **2020**, *121*, 16–28. [[CrossRef](#)]
86. Avikal, S.; Sharma, K.; Barthwal, A.; Kumar, K.N.; Badhotiya, G.K. Estimation of age from speech using excitation source features. *Mater. Today Proc.* **2021**, *46*, 11046–11049. [[CrossRef](#)]
87. Srivastava, R.; Pandey, D. Speech recognition using HMM and Soft Computing. *Mater. Today Proc.* **2022**, *51*, 1878–1883. [[CrossRef](#)]
88. Narendra, N.P.; Alku, P. Automatic intelligibility assessment of dysarthric speech using glottal parameters. *Speech Commun.* **2020**, *123*, 1–9. [[CrossRef](#)]
89. Ilyas, M.; Nait-Ali, A. Auditory perception vs. face based systems for human age estimation in unsupervised environments: From countermeasure to multimodality. *Pattern Recognit. Lett.* **2021**, *142*, 39–45. [[CrossRef](#)]
90. Abirami, B.; Subashini, T.; Mahavaishnavi, V. Automatic age-group estimation from gait energy images. *Mater. Today Proc.* **2020**, *33*, 4646–4649. [[CrossRef](#)]
91. Sethi, D.; Bharti, S.; Prakash, C. A comprehensive survey on gait analysis: History, parameters, approaches, pose estimation, and future work. *Artif. Intell. Med.* **2022**, *129*, 102314. [[CrossRef](#)]
92. Lee, S.; Lee, J.; Moon, H.; Park, C.; Seo, J.; Eo, S.; Koo, S.; Lim, H. A Survey on Evaluation Metrics for Machine Translation. *Mathematics* **2023**, *11*, 1006. [[CrossRef](#)]
93. Aafaq, N.; Mian, A.; Liu, W.; Gilani, S.Z.; Shah, M. Video description: A survey of methods, datasets, and evaluation metrics. *ACM Comput. Surv. (CSUR)*. **2019**, *52*, 1–37. [[CrossRef](#)]
94. Rao, K.S.; Manjunath, K.E. *Speech Recognition Using Articulatory and Excitation Source Features*; Springer International Publishing: Cham, Switzerland, 2017. [[CrossRef](#)]
95. Grzybowska, J.; Kacprzak, S. Speaker Age Classification and Regression Using i-Vectors. In Proceedings of the INTERSPEECH 2016 Conference, San Francisco, CA, USA, 8–12 September 2016; pp. 1402–1406. [[CrossRef](#)]
96. Taheri, S.; Toygar, Ö. On the use of DAG-CNN architecture for age estimation with multi-stage features fusion. *Neurocomputing* **2019**, *329*, 300–310. [[CrossRef](#)]
97. Hiba, S.; Keller, Y. Hierarchical Attention-based Age Estimation and Bias Estimation. *arXiv* **2021**, arXiv:2103.09882.
98. Liu, H.; Lu, J.; Feng, J.; Zhou, J. Label-Sensitive Deep Metric Learning for Facial Age Estimation. *IEEE Trans. Inf. Forensics Secur.* **2018**, *13*, 292–305. [[CrossRef](#)]
99. Hu, Z.; Wen, Y.; Wang, J.; Wang, M.; Hong, R.; Yan, S. Facial Age Estimation With Age Difference. *IEEE Trans. Image Process.* **2016**, *26*, 3087–3097. [[CrossRef](#)]
100. Niu, Z.; Zhou, M.; Wang, L.; Gao, X.; Hua, G. Ordinal Regression with Multiple Output CNN for Age Estimation. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 4920–4928. [[CrossRef](#)]
101. Shin, N.-H.; Lee, S.-H.; Kim, C.-S. Moving Window Regression: A Novel Approach to Ordinal Regression. *arXiv* **2022**, arXiv:2203.13122.

102. Rothe, R.; Timofte, R.; Van Gool, L. Deep Expectation of Real and Apparent Age from a Single Image Without Facial Landmarks. *Int. J. Comput. Vis.* **2016**, *126*, 144–157. [[CrossRef](#)]
103. Wang, X.; Guo, R.; Kambhampettu, C. Deeply-Learned Feature for Age Estimation. In Proceedings of the 2015 IEEE Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 5–9 January 2015; pp. 534–541. [[CrossRef](#)]
104. Duan, M.; Li, K.; Yang, C.; Li, K. A hybrid deep learning CNN-ELM for age and gender classification. *Neurocomputing* **2018**, *275*, 448–461. [[CrossRef](#)]
105. Ng, C.-C.; Cheng, Y.-T.; Hsu, G.-S.; Yap, M.H. Multi-layer age regression for face age estimation. In Proceedings of the 2017 Fifteenth IAPR International Conference on Machine Vision Applications (MVA), Nagoya, Japan, 8–12 May 2017. [[CrossRef](#)]
106. Antipov, G.; Baccouche, M.; Berrani, S.-A.; Dugelay, J.-L. Apparent Age Estimation from Face Images Combining General and Children-Specialized Deep Learning Models. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 801–809. [[CrossRef](#)]
107. Kalluri, S.B.; Vijayasenan, D.; Ganapathy, S. A Deep Neural Network Based End to End Model for Joint Height and Age Estimation from Short Duration Speech. In Proceedings of the ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 6580–6584. [[CrossRef](#)]
108. Singh, J.B.R.; Raj, B. Short-term analysis for estimating physical parameters of speakers. In Proceedings of the 2016 4th International Conference on Biometrics and Forensics (IWBF), Limassol, Cyprus, 3–4 March 2016; pp. 1–6.
109. Garofolo, J.S.; Lamel, L.F.; Fisher, W.M.; Fiscus, J.G.; Pallett, D.S.; Dahlgren, N.L. TIMIT Acoustic-Phonetic Continuous Speech Corpus. Available online: <https://doi.org/10.35111/17gk-bn40> (accessed on 11 January 2023). [[CrossRef](#)]
110. Liu, Y.; Fung, P.; Yang, Y.; Cieri, C.; Huang, S.; Graff, D. HKUST/MTS: A Very Large Scale Mandarin Telephone Speech Corpus. In *Lecture Notes in Computer Science*; Springer: Berlin/Heidelberg, Germany, 2006; pp. 724–735. [[CrossRef](#)]
111. Serda, M. Synteza i aktywność biologiczna nowych analogów tiosemikarbazonowych chelatorów żelaza. *Uniw. Śląski* **2013**, *3*, 343–354.
112. Fung, D.G.P.; Huang, S. HKUST Mandarin Telephone Speech, Part 1-Linguistic Data Consortium. Available online: <https://catalog ldc.upenn.edu/LDC2005S15> (accessed on 20 June 2022).
113. Group, N.M.I. 2008 NIST Speaker Recognition Evaluation Test Set-Linguistic Data Consortium. Available online: <https://catalog ldc.upenn.edu/LDC2011S08> (accessed on 20 June 2022).
114. An, P.; Shenzhen, T. Towards speaker age estimation with label distribution learning. In Proceedings of the ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022; pp. 4618–4622.
115. Ghahremani, P.; Nidadavolu, P.S.; Chen, N.; Villalba, J.; Povey, D.; Khudanpur, S.; Dehak, N. End-to-end Deep Neural Network Age Estimation. *INTERSPEECH* **2018**, *2018*, 277–281. [[CrossRef](#)]
116. Kelly, F.; Drygajlo, A.; Harte, N. Speaker verification with long-term ageing data. In Proceedings of the 2012 5th IAPR International Conference on Biometrics (ICB), New Delhi, India, 29 March–1 April 2012; pp. 478–483. [[CrossRef](#)]
117. Pantraki, E.; Kotropoulos, C. Multi-way regression for age prediction exploiting speech and face image information. In Proceedings of the 2017 25th European Signal Processing Conference (EUSIPCO), Kos, Greece, 28 August–2 September 2017; pp. 2196–2200. [[CrossRef](#)]
118. Kelly, F.; Drygajlo, A.; Harte, N. Speaker verification in score-ageing-quality classification space. *Comput. Speech Lang.* **2013**, *27*, 1068–1084. [[CrossRef](#)]
119. Itou, K.; Yamamoto, M.; Takeda, K.; Takezawa, T.; Matsuoaka, T.; Kobayashi, T.; Shikano, K.; Itahashi, S. JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research. *Acoust. Sci. Technol.* **1999**, *20*, 199–206. [[CrossRef](#)]
120. Kobayashi, T. ASJ Continuous Speech Corpus. *Jpn. Newsp. Artic. Sentences* **1997**, *48*, 888–893.
121. VoxCeleb. Available online: <https://www.robots.ox.ac.uk/~vgg/data/voxceleb/> (accessed on 19 June 2022).
122. Chung, J.S.; Nagrani, A.; Zisserman, A. VoxCeleb2: Deep Speaker Recognition. In Proceedings of the INTERSPEECH 2018, Hyderabad, India, 2–6 September 2018; pp. 1086–1090. [[CrossRef](#)]
123. Nagrani, A.; Chung, J.S.; Zisserman, A.V. VoxCeleb: A large-scale speaker identification dataset. In Proceedings of the Interspeech, Stockholm, Sweden, 20–24 August 2017; pp. 2616–2620.
124. Zhao, M.; Ma, Y.; Liu, M.; Xu, M. The speakin system for voxceleb speaker recognition challenge 2021. *arXiv* **2021**, arXiv:2109.01989.
125. Naohiro, T.V.; Ogawa, A.; Kitagishi, Y.; Kamiyama, H. Age-vox-celeb: Multi-modal corpus for facial and speech estimation. In Proceedings of the ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 6963–6967.
126. Bahari, M.H.; Van Hamme, H. Speaker age estimation using Hidden Markov Model weight supervectors. In Proceedings of the 2012 11th International Conference on Information Science, Signal Processing and their Applications (ISSPA), Montreal, QC, Canada, 2–5 July 2012; pp. 517–521. [[CrossRef](#)]
127. van Leeuwen, D.A.; Kessens, J.; Sanders, E.; Heuvel, H.V.D. Results of the n-best 2008 dutch speech recognition evaluation. *INTERSPEECH* **2009**, *2009*, 2571–2574. [[CrossRef](#)]
128. Spiegl, W.; Stemmer, G.; Lasarczyk, E.; Kolhatkar, V.; Cassidy, A.; Potard, B.; Shum, S.; Song, Y.C.; Xu, P.; Beyerlein, P.; et al. Analyzing features for automatic age estimation on cross-sectional data. In Proceedings of the Tenth Annual Conference of the International Speech Communication Association, Brighton, United Kingdom, 6–10 September 2009. [[CrossRef](#)]
129. Harnsberger, J.D.; Brown, W.S.; Shrivastav, R.; Rothman, H. Noise and Tremor in the Perception of Vocal Aging in Males. *J. Voice* **2010**, *24*, 523–530. [[CrossRef](#)]

130. Burkhardt, F.; Eckert, M.; Johanssen, W.; Stegmann, J. A database of age and gender annotated telephone speech. In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), Valletta, Malta, 17–23 May 2010; pp. 1562–1565.
131. Keren, G.; Schuller, B. Convolutional RNN: An enhanced model for extracting features from sequential data. In Proceedings of the 2016 International Joint Conference on Neural Networks (IJCNN), Vancouver, BC, Canada, 24–29 July 2016; pp. 3412–3419. [[CrossRef](#)]
132. Cao, Y.T.; Iii, H.D. Toward Gender-Inclusive Coreference Resolution. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 4568–4595. [[CrossRef](#)]
133. Cao, Y.T.; Daumé, H. Toward Gender-Inclusive Coreference Resolution: An Analysis of Gender and Bias Throughout the Machine Learning Lifecycle. *Comput. Linguist.* **2021**, *47*, 615–661. [[CrossRef](#)]
134. Bahari, M.H.; McLaren, M.; Van Hamme, H.; van Leeuwen, D.A. Speaker age estimation using i-vectors. *Eng. Appl. Artif. Intell.* **2014**, *34*, 99–108. [[CrossRef](#)]
135. Sadjadi, S.O. NIST SRE CTS Superset: A large-scale dataset for telephony speaker recognition. *arXiv* **2021**, arXiv:2108.07118.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.