Iraqi Journal for Computers and Informatics
Available online: https://ijci.uoitc.edu.iq/index.php/ijci
Volume 52, issue 1. 2026

Print ISSN: 2313-190X, Online ISSN: 2520-4912

From CNNs to Transformers: The Evolution of Neural Architectures in Biometric Fusion Systems – A Narrative Review

Rowida Jasim Alazawi [a], Nada Jasim Habeeb [b], Alaa Jabbar Qasim Almaliki [c]

[a], *Technical College of Management, Middle Technical University, Baghdad, Iraq.*

[b] *Technical College of Management, Middle Technical University, Baghdad, Iraq.*

[c] *School of Computing, Universiti Utara Malaysia (UUM), 06010 Sintok, Kedah Darul Aman, Malaysia.*

ARTICLE INFO

Keywords:

Biometric Fusion, Convolutional Neural Networks (CNNs, Transformers, Multimodal

ABSTRACT

Biometric recognition systems have evolved from uni-modal systems to multi-modal frameworks, improving both accuracy and robustness. Deep learning has been at the forefront of this evolution, and Convolutional Neural Networks CNNs have been the foundation for biometric fusion systems due to their ability to tap the spatial features of the input data. However, CNN architectures have inherent challenges in handling long-distance dependencies, as well as inter-modal dependencies, within the biometric modalities. In this narrative review, the architectural development of deep learning models used in multimodal biometric fusion systems will be critically discussed, from the use of CNN-based models and Recurrent Neural Networks RNNs, including Long Short-Term Memory LSTM architectures, towards the development of hybrid and transformer models. This article will also discuss the different levels of biometric fusion, including fusion at the sensor, feature, score, and decision levels, and will synthesize the challenges associated with multimodal biometric fusion systems. Through the analysis of research gaps in existing studies, as well as the motivations behind the transition to new architectures, this literature review points to the combination of CNNs and Transformers as an area of great promise for the development of scalable, transparent, and robust multimodal biometric fusion systems.

1. INTRODUCTION

Biometric identification is an essential component of document verification and information security systems and is driven by the increasing requirement for safe, convenient, and reliable methods of authentication. Although unimodal biometric systems, such as fingerprint and iris scanning systems, are currently being widely used, there are certain inherent limitations to such systems, such as intra-class variability, inter-class similarities, vulnerability to spoof attacks, and degradation due to uncontrolled environments [1]. However, to overcome the aforementioned drawbacks, there has been an increasing trend of research interest in multimodal biometric identification systems, which use the combination of various biometric characteristics to minimize the rate of errors and maximize the ability of detecting fraud. The advancement of this area has been significantly facilitated by the application of deep learning methodologies, with a focus on Convolutional Neural Networks CNNs. CNN-based approaches have proved to be effective in image-related tasks, ranging from generic computer vision applications [2] to biometric recognition systems. In multimedia biometric systems, which combine modalities like iris, face, and fingerprint, CNNs have been successfully used for spatial extraction and initial fusion [3]. Despite the success that CNN-based models have shown, there exist some inherent structural limitations in these models. This is because the localized receptive fields in these models hinder the

E-mail address:
Dac5008@mtu.edu.iq [a]
Nadaj2013@mtu.edu.iq [b]
alaa.jabbar@uum.edu.my

Corresponding* : Rowida Jasim Alazawi

Received 05th th December 2025, Accepted 27 January 2026

DOI: 10.25195/ijci.v52i1707.

modelling of long-term dependencies and complex inter-modal relations, which are necessary for successful multimodal biometric fusion. With the growing complexity in biometric systems and the diversity in the number of biometric modalities, the above-mentioned limitations become more apparent. Current research shows that the use of transformer models, which incorporate the attention mechanism, is a more efficient approach in modelling global contextual relations in heterogeneous biometric data. This has been evidenced in models that incorporate iris and ECG biometric modes based on the Swin-Transformer architecture, which have shown improved performance over the previous CNN-based models [4]. On the other hand, the need for an effective transition has led to the development of hybrid CNN & Transformer models, which try to combine the strengths of CNNs in local feature extraction with the global modelling power of transformers. These hybrid models have shown great promise in the related area of medical image segmentation, where they successfully combined local and global information [5]. In the area of cliometrics, multi-modal fusion approaches based on CNNs, such as the fusion of finger knuckle print & fingernail patterns at different fusion levels, achieved enhanced accuracy over uni-modal approaches [6]. These methods, however, are still limited by the fact that CNNs are incapable of exploiting the global correlations that are typical of complex biometric patterns. Apart from issues of performance, security and protection of templates have also been treated through CNN-based architectures. In particular, cancelable CNN architectures for the fusion of iris and fingerprint biometric traits have shown remarkably low EERs, which is an indication of high recognition rates and protection of templates against misuse [7]. Although these works prove the success of CNN-based fusion in the initial design phases of biometric solutions, at the same time they also point out the limitations of architectures in terms of scalability and interpretability.

Although there is an increasing number of research works focusing on CNN-based, hybrid, and transformer-based biometric systems, the current reviews have given less emphasis to the architectural evolution that exists between the CNN-based architectures and the attention-based transformer models. This is because there is still minimal research work focusing on the architectural evolution.

The review fills that gap by conducting a critical narrative analysis of the evolution of neural architectures in multi-modal biometric fusion, tracing the path from CNN-based architectures to hybrid architectures and transformers. This review brings focus to architectures, their motivations, trade-offs, and levels of fusion to provide a structured view of open challenges and future research trends.

In keeping with this purpose, the review is informed by the following research questions:

- What were the CNN model architecture limitations that led to the shift to hybrid models and transformer models in multimodal biometric fusion?

- What impact has the development in the architecture of the models, from CNN to Hybrid and then Transform models, had on biometric fusion approaches at various fusion levels, namely Sensor, Feature, Score, and Decision levels?

- How do transformer architectures and the hybrids overcome the issues of modelling global and cross-model relationships as opposed to the previous CNN and RNN_/_LSTM architectures?

- What are the open architectural questions in multimodal biometric fusion systems that remain even with recent advances in transformer architectures?

2. LITERATURE REVIEW METHODOLOGY

This narrative review has been carried out through major academic databases such as Scopus, Web of Science, IEEE Xplore, ScienceDirect, and Google Scholar, by using structurally defined keywords related to multimodal biometrics, biometric fusion, and deep architectures such as CNNs, Transformers, and their combinations. The inclusion criteria have been limited to peer-reviewed journals and high-quality conferences only, with the time span of the inclusion criteria limited to studies published between 2018 and 2025. Studies related to unimodal biometrics, outdated studies, and studies with less technical relevance have not been considered. However, some relevant studies from the early days have been intentionally considered in this narrative review to present theoretical background knowledge related to the architectural advancements towards hybrid and transformer-based multimodal biometric fusion. The selected studies have been reviewed by using a qualitative narrative review process.

3. BASIC CONCEPTS

3.1 Biometric Methods

The identification systems have also incorporated the use of physiological techniques, which include fingerprints, iris scans and even facial expressions as part of the system. These approaches fall under three categories depending on their traits: universality, permanence, and individuality. An example is that fingerprints are formed during pregnancy and assume various patterns, thus creating differences between people, even twins. This biological uniqueness is what makes them a settled reference method of biometric security_[8]. Concerning facial recognition technology, it is true that though extensively used, it has become more

sensitive to external factors like light, orientation, and facial expression, among others. Although recognition accuracy has been greatly improved, mostly due to the development of deep learning, many systematic reviews show that issues still exist, especially in non-controlled settings or in the real world [9]. These limitations were intensified by the COVID-19 pandemic, because the face was largely covered by the mask, which adversely affects the efficacy of the facial recognition systems. Recent research indicates that it is an urgent task to create specialized deep learning architectures. Besides good forgery detection schemes and large and well-classified data used to reduce such impacts [10] [11], and since it has high entropy and stability over time, the iris is one of the most effective and predictive features to be used in the identification of individuals. The conventional system of iris recognition is used to carry out five tasks most commonly, namely, image capture, segmentation, normalization, feature extraction and final matching. Research has shown that segmentation is very important, and therefore, proper segmentation will have a direct influence on recognition. Recent CNNs-based models, and notably U-Net and variants, have been found to be more resistant to frequent obfuscations including eyelids, eyelashes and brilliant reflections [12] [13]. Nevertheless, iris recognition systems are subject to hacking. Counterfeit contact lenses and high-quality reproductions of images are good targets of such serious forgery attacks. Biometric detection mechanisms are used in order to overcome these vulnerabilities. Such techniques have outperformed the traditional, manually developed feature methods in accuracy and generalizability [14]. Despite its strength and popularity, Biometrics has inherent weaknesses. Physiological biometrics have the basic constraints of the environment in which controlled data is gathered, and are prone to falsification. These restrictions have triggered interest in merging behavioral characteristics with multimedia integration models.

Behavioral biometrics is a prominent trend in verification systems, capable of modeling the analysis of behavioral patterns that

individuals naturally display when using smart devices. Recent researches have shown that these patterns are distinguished by a distinct level of individuality and consistency in the systematic recording, which makes them an adversary of physiological measurements [15] [16]. It has been found out that behavioral characteristics like writing, gait and touch offer an extra remedy since it does not involve direct manipulation by the user, as well as being hard to counterfeit [15] [16]. Gait recognition is also one of the most unique behavioral approaches because of its ability to be sampled at long distances, not requiring human cooperation, and relatively stable with respect to its development, as well as its inability to be imitated. Such analyses have been applied to consider feature extraction algorithms, classifier structures, and specialized databases in this area [17]. Other behavioral patterns, including dynamic signatures, voice, and keyboard input dynamics, also have a certain degree of accuracy, albeit they are relatively cheap and easily incorporated into intelligent verification systems [18].

Despite the challenges associated with behavioral change over time and its susceptibility to environmental conditions, the general trend in research indicates a significant performance improvement when multiple behavioral patterns are integrated or when deep learning techniques are used to address variability and enhance the generalizability of models.

3.2 Levels of Biometric Fusion

Integration methods in multimodal biometric systems are divided into four main levels: sensor, feature, score, and decisio. Each level has unique advantages but also inherent technical challenges in processing and integration. Figure 1 illustrates the different levels of multimodal biometric fusion.

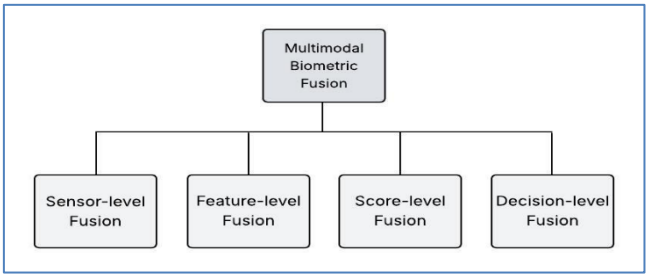

Figure 1: *Different levels of multimodal biometric fusion*

3.2.1 Sensor-Level Fusion

At this stage, a combination of biometric data of various sensors is done and then features are extracted. This gives a full representation of the biometric information, though it also needs a high accuracy of synchronization and calibration between sensors. Indicatively, multispectral facial recognition studies have revealed that when only visible light is used, this will greatly decrease the accuracy of uncontrolled data. Conversely, the solution to the weak point of the system is to add near-infrared/thermal imaging channels [19] [20]. Another technology that has been offered to combine high and low-frequency

details is microwave-based pixel fusion among others and enhancing cross-spectral recognition [21] [22]. Sensor-level integration has the following benefits, but is computationally intensive and usually not practical in a real-time application since it demands a precise sensor alignment.

3.2.2 Feature-Level Fusion

Feature-level data integration is a method of combining vectors derived from different methods or techniques into a unified representation. This type of integration contains a vast amount of information, allowing the system to exploit correlations and integrations between different techniques, thereby improving the accuracy of analysis or prediction. Among the most commonly used methods are vector integration, dimensionality reduction, and attention-based weighting mechanisms. For example, integrating fingerprint and signature features early can achieve higher accuracy than integrating them later [23]. Furthermore, facial and finger vein recognition has been shown to reach up to 98% accuracy using an attention-based integration model [24]. However, template security remains a significant concern. To address this issue, newer methods, such as reversible softmaxOut integration networks, have been developed to maintain high levels of recognition accuracy while ensuring the protection of sensitive biometric information [25].

3.2.3 Score-Level Fusion

Score-level merging is one of the most common methods because it strikes a good balance between ease of use and accuracy. Instead of using raw features, we normalize and merge matching scores with each classifier using methods such as reference summation or logistic regression [26]. Notable examples include the merging of 3D facial and ear recognition systems, which achieved 99.25% accuracy [27] , and the merging of finger and vein texture features, which achieved 99.62% accuracy [28].

However, score-level merging methods may not make optimal use of discriminatory information compared to feature-level merging because they operate on smaller intermediate outputs.

3.2.4 Decision-Level Fusion

Decision integration in multi-biometric systems is implemented by combining independent decisions from each module after they have been separately classified. A unified decision is reached by using AND / OR rules or more advanced techniques such as Majority voting, Bayesian fusion, and Dempster-Shafer. This increases verification reliability when only binary decisions from matching systems are in agreement [29]. Modern banking applications have demonstrated that probabilistic decision techniques, such as pattern integration via Dempster-Shafer theory, give the system a greater ability to balance rejecting a legitimate user and preventing the acceptance of a forger, especially when there is conflict or uncertainty between biometric patterns [30]. Furthermore, literature reviews indicate that this level is most effective when used as a final layer on top of information-rich integration levels, such as feature or score integration, to reach a decision in highly sensitive multi-biometric systems [31]. Table 1 summarizes the comparison of recent multimodal biometric fusion approaches.

Table 1: *Comparative summary of recent multimodal biometric fusion studies*

Ref	Modalities	Fusion Level	Dataset Scale	Model Type	Key Performance Metric
[32]	Face + Fingerprint	Score-level	150 subjects (450 face + 450 FP images)	Wiener filter + BRO-DCNN	Accuracy= 98 %, Precision =99 %, Specificity=96 %, Sensitivity =94 %
[33]	Face + Fingerprint	Feature-level	SDUMLA-HMT (~900 subjects)	SIFT + BCO-AKSVM	f1-score=99 %, recall=98 %, accuracy=97 %, precision=88 %
[34]	Latent Fingerprint + Iris	Score-level	Public benchmark datasets	CNN + Gabor matching	EER values as low as 0.043
[35]	Fingerprint + Face + Iris + DNA	Score-level	500 subjects (140 tetra-modal)	Modality-specific matchers + Choquet Integral (PSO)	EER = 0.1%

[36]	Face + Fingerprint + Palmprint + Voice + ECG + Ear + Periocular	Score- & Decision-level	306 subjects (9 modalities, age 8–90)	CNNs + RNNs (baseline evaluation)	Accuracy = 0.998
[37]	Face + Fingerprint	Score-level (spoof-aware)	Public anti-spoofing benchmarks	Anomaly-based fusion + bagging	EER = 0.47–1.81%
[38]	Face + Dynamic Signature	Feature-level	25 subjects (1750 samples)	CNN + LSTM/GRU/TCN	Accuracy=98%, f1-score= 97.9%

As shown in table 1, the performance differences are significant with respect to the fusion level and the size of the dataset. Smaller and more controlled studies are likely to report more optimistic measures of accuracy, while more realistic measures, such as Equal Error Rate, are more typical of large-scale tests of security-related benchmarks. On the other hand, more recent work has started using deep learning models for combining behavior and physiological biometrics, while dataset-related studies are providing necessary baselines.

4. NEURAL ARCHITECTURES

Neural network architecture is an essential component of AI research and has continued to develop as the complexity of the data being handled has increased. This has significantly improved the efficiency of models in learning and processing data. It is evident that the development is moving away from the traditional models and towards advanced models based on attention mechanisms.

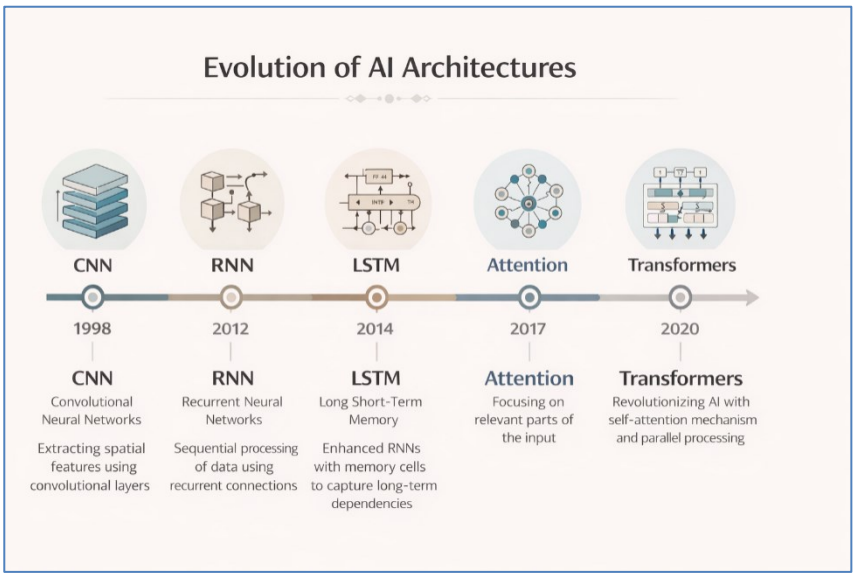

Figure 2: *The evolution of AI architectures from CNN to Transformers*

4.1 Convolutional Neural Networks CNNs

CNNs are a fundamental architecture in deep learning, particularly in computer vision. Their advantage lies in their ability to automatically learn a set of features and extract deep feature representations from raw visual data. This makes them effective in single-mode biometric tasks such as iris recognition, facial recognition, and fingerprint matching. As research has progressed, more complex CNN architectures have been developed, such as DenseNet, VGG and ResNet. These models have managed to overcome the main structural issues like gradient vanishing and have also added new features like residual connections. Through these architectural enhancements, the networks have been able to become deeper, more stable, and more efficient [39] [40] [41] [42]. Although CNNs are very efficient in deriving spatial patterns, they are not easy to use. They involve very complicated computational processing and massive and classified training datasets, which are not feasible when the resources are limited or when there is an immediate requirement in the application. CNNs have a weakness of not extracting long-range relationships and cannot combine features of different sources, like images and sound, so they cannot be easily applied in multimodal biometric systems. In order to address such constraints, hybrid models are being developed in order to fill the gap between CNNs and Transformer models. These hybrids make use of the advanced spatial analysis of convolutional networks and the temporal or cross-modal context of Transformer models [41].

4.2 RNNs and LSTMs

RNNs are made in a specific way to process sequential data. Such networks also employ some form of memory, which enables them to store the information remembered during the past, and it is therefore an appropriate type of network to analyze behavioral biometrics, including gait patterns, voice patterns and keystroke patterns [43] [44]. Yet, the classical RNNs have a disadvantage of gradient vanishing, thus they cannot easily know how data relates with time. Long Short-Term Memory LSTMs and Gated Recurrent Units GRUs have been able to solve this problem. This network architecture exploits advanced memory through smart gates, and it is more effective in governing the information flow over time, thereby assisting the model in short term retention of useful intermediary results [45] [46]. The uses of LSTMs are not limited to biometrics; they are also used in fields such as hydrology, natural language processing, sensor-based prediction systems, and other advanced applications [47] [48] [49] [50] [51]. However, these networks may overburden computers and do not operate as efficiently as newer models. Attention-based models are gradually gaining a larger share of Long Short-Term Memory Networks LSTMs in integrating behavioral biometric data; however, LSTMs are still useful.

4.3 Transformers and Attention Mechanisms

Attention mechanisms have fundamentally changed how neural networks handle data. Initially introduced for tasks such as machine translation [52]. Attention enables models to focus on the most important parts of an input sequence at any given moment. Transformers have further developed this mechanism by replacing the sequential nature of Recurrent Neural Networks RNNs with self-attention layers. This new approach allows for simultaneous data processing and a better understanding of long-range dependencies, significantly accelerating tasks involving large datasets [53]. In computer vision, Vision Transformers ViTs have adopted this concept by segmenting images and processing them as sequences. The result is superior performance in classification and recognition [54] [55]. Transformers have also been effectively used in biometrics as receiver models or integrated with CNNs. in hybrid systems. These methods have achieved remarkable results in integrating multimodal biometric data, combining physiological (fingerprints, face, veins) and behavioral (gait, voice) identifiers [56] [57]. Another key advantage of attention mechanisms is their transparency, as they provide the possibility of interpretation by highlighting the parts of the input that the model focuses on for decision-making, a crucial feature for biometric systems that require trust and accountability [58] [59]. However, transformers models still consume significant resources despite all the advancements. The main characteristics and performance differences of common deep learning models are presented in table 2

Table 2: *Comparison of Deep Learning Architectures in Terms of Structure, Uses, and Performance*

Ref	Structure	Uses	Performance
[60] [61]	CNNs (Conv and Pooling)	Images and computer vision	Strong in vision but weak in sequence
[62] [63]	RNN	Text, speech, and time series	Good for short, but weak for long sequences.
[64] [51]	LSTM (RNN and Gates)	Long texts and translation	For long sequences, it's better than RNN.
[52] [58]	Attention	NLP and machine translation	Solves context-loss problem
[53] [65]	Transformers	NLP, vision, multimodal	The strongest and fastest architecture

5. HISTORICAL DEVELOPMENT OF NEURAL STRUCTURES

5.1 Convolutional Neural Networks and Feature-Level Integration

CNNs have the capability of automatically deriving hierarchical and complicated features of the raw data. This has changed the multimodal biometric data integration methodology. CNNs are a major advancement in the domain of feature-level integration processes because they have the capacity to combine various biometric inputs, e.g. facial characteristics, iris scans, and fingerprints, in one image. This integration enhances the accuracy of the system in addition to making the system more efficient, reliable and less repetitive. There is an emerging evidence on the improvement of biometric authentication systems that shows that the method is much better in enhancing the overall performance of the systems. Indicatively, a research study [66] on multimodal hand biometrics revealed that the feature integration of deep learning was significantly better than the conventional, manually designed feature methods. A second example is an investigation [67] which created a convolutional neural network-based model with the help of face, finger veins, and iris features. Their incorporation technique led to an almost perfect recognition and ability, obviously exceeding the capacity of single mode systems. Likewise, [68] examined a fingerprint and ECG hybrid type of data by sequential integration approach with CNNs. This compound was not only able to capture the features of space, but it was also able to encode physiological signals, and thus identity verification was more accurate. Most intriguingly, data integration with CNNs is also effective outside the conventional biometrics, as well. An example is the work of [69] who created an audiovisual network of data which fused facial and speech. Their model was also better than single-mode systems in voice verification and identity. Though not directly correlated with biometrics, CNNs integration is among the most effective

and new approaches to the combination of various biometric patterns. Moreover, this practice has certain disadvantages. CNNs are not appropriate with temporal and sequential data since they demand a lot of computing power and cannot be used in gait patterns or keystroke patterns. Thus, they are not that useful in certain cases. Those restrictions shed light on the increased demand of models that can effectively work with sequence dynamics, such as recurrent networks or attention-based models, which will be discussed in the subsequent sections.

5.2 Fusion Using LSTMs and RNNs

CNNs are ideal for biometric systems. However, they are not well-suited for modelling time-changing patterns. This limitation is particularly pronounced in dynamic or sequential biometric patterns, such as Electrocardiogram ECG signals, dynamic fingerprints, and gait analysis. To overcome this shortcoming, Recurrent Neural Networks RNNs have emerged as a promising technological solution, but Long Short-Term Memory networks LSTMs have proven more successful. LSTMs are characterized by an internal memory state and "gate" control, meaning they can retain information for extended periods, unlike conventional recurrent networks. These characteristics make them particularly suitable for handling time-changing biometric data and are therefore well-suited for analyzing this type of data. Consider, for example, ECG-based biometric authentication. A deep LSTM was applied to model the sequential nature of cardiac signals, resulting in highly accurate identification. This demonstrates that adding temporal modelling to recognition tasks significantly improves performance, [70]. Hybrid Convolutional Neural Network CNNs-LSTM architectures are gaining popularity for improving biometric data integration. In these models, CNN layers are first used to detect spatial features, and then LSTM layers are used to learn how the sequence depends on time. This layered approach enables feature-level integration, combining the spatial sensitivity of CNNs with the time-learning ability of LSTMs [71]. An example of a hybrid approach in the real world is multimodal biometric systems. Here, CNN layers were applied to obtain initial features, and LSTM layers were applied to find out how things evolve with time. This approach enhanced the accuracy and the stability of the recognition process, particularly when multiple biometric features are combined [72]. LSTMs are also very flexible in terms of incorporating multimodal biometric information, e.g. the integration of facial image with dynamic input. In one study, TCN, GRU, LSTM and CNN networks were used together to align spatial facial features to the temporal information presented by handwritten signatures. Among these networks that have been the most important in terms of modelling sequential signature patterns and matching them with stationary face information are the Long Short-Term Memory networks LSTMs. The incorporation of this fusion method is strong and effective in different situations of data collection [38]. Besides, LSTM models also provide a solution to continuous authentication systems, where a user is continually authenticated according to the current receipt of behavioral and physiological samples. In short, the combination of Recurrent Neural Networks RNNs and Long Short-Term Memory networks LSTMs into biometric data fusion schemes allows designing intelligent recognition systems which are capable of adapting to the environment in which they are deployed and knowing which task they are supposed to accomplish, which is paramount in the current state of cybersecurity [73]. In brief, the combination of Recurrent Neural Networks RNNs and Long Short-Term Memory networks LSTMs in the biometric data fusion schemes enables achieving a quantum leap in the design of intelligent recognition systems, which are able to adapt to the spatial and temporal continuity is also done exceptionally well in this architecture, opening the way to even more developments like transformer-based fusion models and attentional mechanisms [38] [74].

5.3 Attention Mechanisms and Transformer-Based Fusion in Biometrics

Attention mechanisms, particularly those popularized by transformer architectures, are essential for enhancing the capabilities of biometric integration systems. As previously mentioned, these mechanisms were originally developed for natural language processing models and help the model focus on the most important inputs. In the context of biometrics, this means that the model can give greater weight to high-quality data and less weight to noisy or poorly informed data. The result? Less random and more discriminating and stable representations. Recent research highlights the impact of attention mechanisms during the integration phase of multimodal biometrics. As a practical example, researchers at [75] developed a multimodal fingerprint recognition model by combining two mechanisms: channel focus and a cryptographic integration strategy. This approach helps the model find more useful channels in different types of media, such as finger veins and fingerprints. This reduces unnecessary redundancy and improves recognition accuracy by preserving the structural relationships between these media types. At the same time, he was among the few who discovered different Transformer architectures for mobile device gait recognition[76]. These architectures include Vanilla Transformer Informer, Autoformer and Block- Recurrent Transformer. They found that Transformers performed better than standard CNNs and RNN models at least on the ability to capture time dependencies over long times in devices or environments of limited resources. To elaborate on the area of foot gait recognition,[77] trained a self-learning model based on a variety of variants of the Vision Transformer (ViT), including TwinsSVT and CrossFormer. Their model had the ability to learn the dynamics of motion without being provided with labeled data. It was also interesting to note that the hierarchical attention of CrossFormer made it experience high performance even in the case of a noisy real-world gait data. As an example, [78] applied a ViT-based model referred to as Gait-ViT with gait energy images (GEIs). This data was divided into small parts and the model obtained almost perfect results in a number of tests. This showed that attention-based models can be effective in managing variations in gait patterns and viewpoints. Along with the gait analysis, attention

mechanisms are also employed in physiological biometrics. In their study,[79] introduced the FV-MViT which is a lightweight Transformer based finger vein recognition. Their model employed a self-attention that was decoupled and using a two-path residual block to effectively extract both local and global information. The net effect of this was a high degree of accuracy and very little power consumed in processing. In a different paper,[80] proposed the AuthFormer identity verification framework, which was a Transformer-based framework that could be modified to suit the requirements of older users. The system integrates all available biometric signals with dynamism which involve tabulated residual networks and cross-attention modules. It had a high level of accuracy of almost 99 percent and it had a low level of complexity which is a very important aspect since it is easy to use in practical use.

Attention mechanisms, in their various forms—from self-attention to channel focusing and multimodal integration—shape biometric data integration methods. Integrating these mechanisms into the Transformer architecture has been pivotal in developing smarter, more flexible, and scalable systems that clearly outperform traditional methods in terms of accuracy and efficiency.

6. A COMPARISON OF NEURAL ARCHITECTURES IN BIOMETRIC FUSION SYSTEMS

A convolutional neural network (CNN) has always formed the foundation of biometric recognition systems because it is good at local patterns, edges, and the detection of spatial features such as texture. Nevertheless, such networks have intrinsic receptive field limitations, which limit the capacity of their networks to pick up global contextual associations, which is of key importance in complex biometric tasks that require positional effects, occlusions, and multimodal inputs. This weakness has been validated by recent literature, which found that CNNs can reliably and quickly represent local representations, but fail to represent long-range relationships[81]. Conversely, transformer-based architectures can represent the global relationship between spatial (or sequential) data using attention mechanisms. These models are more suited to handle long-range dependencies, and this is why they perform better than CNNs in tasks such as face recognition in challenging scenarios, such as when the distance varies, and an object gets in the line of sight or the lighting changes. [81]. More and more people are moving away from using only one architecture and toward using hybrid models. Hybrid CNN-RNN architectures, like CNN-LSTM, find a good balance between extracting spatial features and modelling time. For instance, a study by[82] found that combining CNN with LSTM got almost perfect accuracy in finding deepfakes by capturing both spatial texture and temporal dynamics. In the same way, [83] combined BiLSTM with a self-attention mechanism after using CNNs to extract features into a multimodal biometric system, and got an accuracy of 98.5%, which was better than the accuracy of each model on its own. The TransUMobileNet model, which uses CNN encoders and transformer decoders, was created by [84] on the transformer-based side. It is highly accurate (95.6%) in medical image segmentation. This is achieved by the use of CNNs that search for patterns in small regions (local search) and transformers, which show how complex global relationships work. These comparative results represent a conceptual shift from shallow, task-oriented architectures to more flexible systems that are aware of the overall context. Although CNNs remain effective in lightweight applications, the harmonious integration of CNNs and Transformers appears to be the most promising path, especially in biometric fusion systems that require high accuracy and advanced contextual depth.

7. CHALLENGES AND RESEARCH GAPS IN DEEP LEARNING-BASED BIOMETRIC VERIFICATION SYSTEMS

7.1 Challenges

- Though there has been great progress achieved in deep learning-based biometric verification systems, some issues still exist, which impede the application of multimodal systems. The first issue is the small amount of available biometric data, as well as its diversity and time span.

- Moreover, the recent architecture of deep learning models has imposed high computational and memory complexities. Another important challenge is the interpretability of deep models. It has been seen that high recognition accuracy is obtained at the cost of transparency and trust in deep models.

- From a security perspective, the resistance of biometric systems to spoofing and adversarial attacks is an area which has yet to be adequately explored. Finally, biometric verification models are often created as separate modules, without being extensively integrated into the overall cybersecurity setup.

7.2 Research Gaps

- This review lists some gaps in the literature that hamper the progress of multimodal biometric verification systems. There is a glaring need for multimodal databases that can be used for the long-term assessment of the system's performance.

- Another major gap is the lack of focus on resource-efficient architectures that consider the trade-off between accuracy and the feasibility of implementation. Furthermore, the incorporation of interpretability into biometric fusion systems is not well explored, although this is vital for trust and security.

- Lack of standardized approaches to robustness assessment in adversarial scenarios is another gap. Lastly, there is a need to develop comprehensive, security-focused approaches to integrate multimodal biometric verification within the existing cybersecurity paradigm, as opposed to viewing biometric verification as a mere recognition problem.

8. FUTURE DIRECTIONS

The next generation of biometric fusion systems will depend on the use of novel neural architectures able to efficiently process and combine multimodal stimuli.

Scalable: Generating models which is required to successfully cope with large sets of biometric data in a fast and efficient way.

Real-time transaction: The time (speed / challenge response) in biometric systems is a critical requirement in order to have truly real-time identification and verification, specifically in those applications that require a quick yet safe response.

Robustness: Algorithms need to remain robust against common natural variations in biometric data (e.g., changes in lighting, light vs darkness, posture and ageing of the subject) while exhibiting the same level of accuracy and stability.

These emerging biometric fusion systems are able to derive better accuracy, reliability and generalization across multiple scenarios by addressing these challenges.

9. CONCLUSION

This review presents how quickly neural architectures have changed in biometric fusion systems. For example, CNNs are used for image processing, while transformers are a new way to model long-range dependencies and combine data from different sources in an efficient way. These architectures have gradually improved systems, which have become both more accurate and easier for system designers to configure, as long as they have overcome the problems in previous versions. And even with these successes, there are still problems to solve. For instance, there are still unresolved problems concerning learning decisions, handling attacks and addressing data and computational requirements. People seem to think that those, paired with some of these newer technologies," graph networks and distributed learning," could make biometric systems a whole lot safer and more useful. The aim for the future is hybrid or balanced models that perform well, are interpretable and will keep people safe in smart cybersecurity environments.

REFERENCES

[1] U. Sumalatha, K. K. Prakasha, S. Prabhu, and V. C. Nayak, "A Comprehensive Review of Unimodal and Multimodal Fingerprint Biometric Authentication Systems: Fusion, Attacks, and Template Protection," IEEE Access, vol. 12, no. April, pp. 64300–64334, 2024, doi: 10.1109 / ACCESS.2024.3395417.

[2] J. Maurício, I. Domingues, and J. Bernardino, "Comparing Vision Transformers and Convolutional Neural Networks for Image Classification: A Literature Review," *Appl. Sci.*, vol. 13, no. 9, 2023, doi: 10.3390 / app13095521.

[3] S. Soleymani, A. Dabouei, H. Kazemi, J. Dawson, and N. M. Nasrabadi, "Multi-Level Feature Abstraction from Convolutional Neural Networks for Multimodal Biometric Identification," *Proc. - Int. Conf. Pattern Recognit.*, vol. 2018-Augus, no. i, pp. 3469–3476, 2018, doi: 10.1109 / ICPR.2018.8545061.

[4] R. Garg, P. Pathak, and M. P. Singh, "A multimodal biometric recognition system based on Fingerprints, Iris and ECG via Swin Transformer and CNN Model," *Syst. Soft Comput.*, vol. 7, no. July, p. 200369, 2025, doi: 10.1016/j.sasc.2025.200369.

[5] D. Sotoude, M. Hoseinkhani, and A. Amiri Tehranizadeh, "Context-aware fusion of transformers and CNNs for medical image segmentation," *Informatics Med. Unlocked*, vol. 43, no. November, p. 101396, 2023, doi: 10.1016/j.imu.2023.101396.

[6] H. Heidari and A. Chalechale, "Biometric authentication using a deep learning approach based on different level fusion of finger knuckle print and fingernail," *Expert Syst. Appl.*, vol. 191, no. July 2020, p. 116278, 2022, doi: 10.1016/j.eswa.2021.116278.

[7] D. K. Vallabhadas, M. Sandhya, S. D. Reddy, and D. Satwika, "Biomedical Signal Processing and Control Biometric template protection based on a cancelable convolutional neural network over iris and fingerprint," *Biomed. Signal*

Process. Control, vol. 91, no. January, p. 106006, 2024, doi: 10.1016/j.bspc.2024.106006.

[8] W. Yang, S. Wang, N. M. Sahri, N. M. Karie, M. Ahmed, and C. Valli, "Biometrics for internet-of-things security: A review," *Sensors*, vol. 21, no. 18, pp. 1–26, 2021, doi: 10.3390/s21186163.

[9] M. K. Hasan, M. S. Ahsan, Abdullah-Al-Mamun, S. H. S. Newaz, and G. M. Lee, "Human face detection techniques: A comprehensive review and future research directions," *Electron.*, vol. 10, no. 19, 2021, doi: 10.3390/electronics10192354.

[10] A. Alzu'bi, F. Albalas, T. Al-Hadhrami, L. B. Younis, and A. Bashayreh, "Masked face recognition using deep learning: A review," *Electronics*, vol. 10, no. 21, p. 2666, 2021.

[11] M. Mahmoud, M. S. E. Kasem, and H. S. Kang, "A Comprehensive Survey of Masked Faces: Recognition, Detection, and Unmasking," *Appl. Sci.*, vol. 14, no. 19, 2024, doi: 10.3390/app14198781.

[12] J. R. Malgheet, N. B. Manshor, and L. S. Affendey, "Iris Recognition Development Techniques: A Comprehensive Review," *Complexity*, vol. 2021, 2021, doi: 10.1155/2021/6641247.

[13] M. R. Sumi, P. Das, A. Hossain, S. Dey, and S. Schuckers, "A Comprehensive Evaluation of Iris Segmentation on Benchmarking Datasets," *Sensors*, vol. 24, no. 21, pp. 1–19, 2024, doi: 10.3390/s24217079.

[14] K. Smita, S. Ahirrao, S. Phansalkar, K. Kotecha, S. Gite, and S. D. Thepade, "Iris Liveness Detection for Biometric Authentication :," pp. 1–54, 2021.

[15] I. Stylios, S. Kokolakis, O. Thanou, and S. Chatzis, "Behavioral biometrics & continuous user authentication on mobile devices : A survey," *Inf. Fusion*, vol. 66, no. February 2020, pp. 76–99, 2021, doi: 10.1016/j.inffus.2020.08.021.

[16] A. Mahfouz, T. M. Mahmoud, and A. Sharaf, "Journal of Information Security and Applications A survey on behavioral biometric authentication on smartphones," *J. Inf. Secur. Appl.*, vol. 37, pp. 28–37, 2017, doi: 10.1016/j.jisa.2017.10.002.

[17] J. Pal, S. Sanjeev, J. Sakshi, A. Uday, and P. Singh, *A Survey of Behavioral Biometric Gait Recognition : Current Success and Future Perspectives*, no. 0123456789. Springer Netherlands, 2019. doi: 10.1007/s11831-019-09375-3.

[18] S. Adnan and B. Alhayani, "Materials Today : Proceedings A comprehensive survey on the biometric systems based on physiological and behavioural characteristics," *Mater. Today Proc.*, vol. 80, pp. 2642–2646, 2023, doi: 10.1016/j.matpr.2021.07.005.

[19] P. Martins, J. S. Silva, and A. Bernardino, "Multispectral Facial Recognition in the Wild," *Sensors*, vol. 22, no. 11, 2022, doi: 10.3390/s22114219.

[20] "Using infrared to improve face recognition of individuals with highly pigmented skin," *iScience*, vol. 26, no. 7, p. 107039, 2023, doi: 10.1016/j.isci.2023.107039.

[21] R. Lionnie, J. Andika, and M. Alaydrus, "A New Approach to Recognize Faces Amidst Challenges: Fusion Between the Opposite Frequencies of the Multi-Resolution Features," *Algorithms*, vol. 17, no. 11, 2024, doi: 10.3390/a17110529.

[22] Z. Cao, N. A. Schmid, S. Cao, and L. Pang, "GMLM-CNN: A Hybrid Solution to SWIR-VIS Face Verification with Limited Imagery," *Sensors*, vol. 22, no. 23, pp. 1–20, 2022, doi: 10.3390/s22239500.

[23] M. Leghari, S. Memon, L. Das Dhomeja, A. H. Jalbani, and A. Ali Chandio, "Deep feature fusion of fingerprint and online signature for multimodal biometrics," *Computers*, vol. 10, no. 2, pp. 1–15, 2021, doi: 10.3390/computers10020021.

[24] Y. Wang, D. Shi, and W. Zhou, "Convolutional Neural Network Approach Based on Multimodal Biometric System with Fusion of Face and Finger Vein Features," *Sensors*, vol. 22, no. 16, pp. 1–15, 2022, doi: 10.3390/s22166039.

[25] J. Kim and Y. G. Jung, "applied sciences Multimodal Biometric Template Protection Based on a Cancelable SoftmaxOut Fusion Network," 2023.

[26] S. Concas *et al.*, "Analysis of Score-Level Fusion Rules for Deepfake Detection," *Appl. Sci.*, vol. 12, no. 15, pp. 1–21, 2022, doi: 10.3390/app12157365.

[27] S. Tharewal *et al.*, "Score-Level Fusion of 3D Face and 3D Ear for Multimodal Biometric Human Recognition," *Comput. Intell. Neurosci.*, vol. 2022, 2022, doi: 10.1155/2022/3019194.

[28] S. A. Haider *et al.*, "An Improved Multimodal Biometric Identification System Employing Score-Level Fuzzification of Finger Texture and Finger Vein Biometrics," *Sensors*, vol. 23, no. 24, 2023, doi: 10.3390/s23249706.

[29] S. Kumar, S. Modak, and V. K. Jha, "Multibiometric fusion strategy and its applications : A review," *Inf. Fusion*, vol.

49, no. January 2018, pp. 174–204, 2019, doi: 10.1016/j.inffus.2018.11.018.

[30] P. Szczuko, A. Harasimiuk, and A. Czyżewski, "Evaluation of Decision Fusion Methods for Multimodal Biometrics in the Banking Application," *Sensors*, vol. 22, no. 6, 2022, doi: 10.3390/s22062356.

[31] M. N. Nachappa, "A Review on Various Fusion Techniques in Multimodal Biometrics," vol. 4, no. 21, pp. 1–8, 2016.

[32] J. Bhuvana, A. Barve, S. Pradeep, and S. Dikshit, "Measurement : Sensors Image sensor fusion for multimodal biometric recognition in mobile devices," *Meas. Sensors*, vol. 36, no. August 2023, p. 101309, 2024, doi: 10.1016/j.measen.2024.101309.

[33] V. Vekariya, M. Joshi, S. Dikshit, and S. K. Manju, "Measurement : Sensors Multi-biometric fusion for enhanced human authentication in information security," *Meas. Sensors*, vol. 31, no. August 2023, p. 100973, 2024, doi: 10.1016/j.measen.2023.100973.

[34] S. Shreya and K. Chatterjee, "Latent fingerprint and Iris fusion for enhancement of performance of human identification system," *Expert Syst. Appl.*, vol. 235, no. August 2023, p. 121208, 2024, doi: 10.1016/j.eswa.2023.121208.

[35] L. R. Haddada *et al.*, "A benchmark tetra-modal biometric score database," *Biomed. Signal Process. Control*, vol. 98, p. 106778, 2024.

[36] R. Yang, Q. Zhang, L. Meng, C. Wang, and Y. Hu, "LUTBIO : A Comprehensive multimodal biometric database targeting middle-aged and elderly populations for enhanced identity authentication ☆," *Inf. Fusion*, vol. 118, no. July 2024, p. 102945, 2025, doi: 10.1016/j.inffus.2025.102945.

[37] P. Wild, P. Radu, L. Chen, and J. Ferryman, "Robust multimodal face and fi ngerprint fusion in the presence of spoo fi ng attacks," *Pattern Recognit.*, vol. 50, pp. 17–25, 2016, doi: 10.1016/j.patcog.2015.08.007.

[38] S. Salturk and N. Kahraman, "Deep learning-powered multimodal biometric authentication: integrating dynamic signatures and facial data for enhanced online security," *Neural Comput. Appl.*, vol. 36, no. 19, pp. 11311–11322, 2024, doi: 10.1007/s00521-024-09690-2.

[39] L. Zewen, L. Fan, Y. Wenjie, P. Shouheng, and Z. Jun, "A survey of convolutional neural networks: analysis, applications, and prospects," *IEEE Trans. neural networks Learn. Syst.*, vol. 33, no. 12, pp. 6999–7019, 2021.

[40] M. Krichen, "Convolutional neural networks: A survey," *Computers*, vol. 12, no. 8, p. 151, 2023.

[41] I. D. Mienye and T. G. Swart, "A comprehensive review of deep learning: Architectures, recent advances, and applications," *Information*, vol. 15, no. 12, p. 755, 2024.

[42] I. D. Mienye, T. G. Swart, G. Obaido, M. Jordan, and P. Ilono, "Deep Convolutional Neural Networks in Medical Image Analysis: A Review," *Inf.*, vol. 16, no. 3, pp. 1–28, 2025, doi: 10.3390/info16030195.

[43] V. S. Lalapura, V. R. Bhimavarapu, J. Amudha, and H. S. Satheesh, "A Systematic Evaluation of Recurrent Neural Network Models for Edge Intelligence and Human Activity Recognition Applications," *Algorithms*, vol. 17, no. 3, 2024, doi: 10.3390/a17030104.

[44] F. H. Quradaa, S. Shahzad, and R. S. Almoqbily, *A systematic literature review on the applications of recurrent neural networks in code clone research*, vol. 19, no. 2 February. 2024. doi: 10.1371/journal.pone.0296858.

[45] F. Rivas, J. E. Sierra-Garcia, and J. M. Camara, "Comparison of LSTM- and GRU-Type RNN Networks for Attention and Meditation Prediction on Raw EEG Data from Low-Cost Headsets," *Electron.*, vol. 14, no. 4, pp. 1–33, 2025, doi: 10.3390/electronics14040707.

[46] M. Waqas and U. W. Humphries, "A critical review of RNN and LSTM variants in hydrological time series predictions," *MethodsX*, vol. 13, no. July, p. 102946, 2024, doi: 10.1016/j.mex.2024.102946.

[47] A. Sherstinsky, "Fundamentals of Recurrent Neural Network RNN and Long Short-Term Memory LSTM network," *Phys. D Nonlinear Phenom.*, vol. 404, no. March, pp. 1–43, 2020, doi: 10.1016/j.physd.2019.132306.

[48] I. Malashin, V. Tynchenko, A. Gantimurov, V. Nelyub, and A. Borodulin, "Applications of Long Short-Term Memory LSTM Networks in Polymeric Sciences: A Review," *Polymers (Basel).*, vol. 16, no. 18, pp. 1–44, 2024, doi: 10.3390/polym16182607.

[49] C. Fu, C. Gao, and W. Zhang, "RUL Prediction for Piezoelectric Vibration Sensors Based on Digital-Twin and LSTM Network," *Mathematics*, vol. 12, no. 8, 2024, doi: 10.3390/math12081229.

[50] F. Kratzert, M. Gauch, D. Klotz, and G. Nearing, "HESS Opinions: Never train a Long Short-Term Memory LSTM network on a single basin," *Hydrol. Earth Syst. Sci.*, vol. 28, no. 17, pp. 4187–4201, 2024, doi: 10.5194/hess-28-4187-

2024.

[51] K. Greff, R. K. Srivastava, J. Koutnik, B. R. Steunebrink, and J. Schmidhuber, "LSTM: A Search Space Odyssey," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 28, no. 10, pp. 2222–2232, 2017, doi: 10.1109/TNNLS.2016.2582924.

[52] D. Bahdanau, K. H. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc.*, pp. 1–15, 2015.

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

[54] A. Dosovitskiy *et al.*, "an Image Is Worth 16X16 Words: Transformers for Image Recognition At Scale," *ICLR 2021 - 9th Int. Conf. Learn. Represent.*, 2021.

[55] M. Rodrigo, C. Cuevas, and N. García, "Comprehensive comparison between vision transformers and convolutional neural networks for face recognition tasks," *Sci. Rep.*, vol. 14, no. 1, pp. 1–10, 2024, doi: 10.1038/s41598-024-72254-w.

[56] R. Garcia-Martin and R. Sanchez-Reillo, "Vision Transformers for Vein Biometric Recognition," *IEEE Access*, vol. 11, no. October 2020, pp. 22060–22080, 2023, doi: 10.1109/ACCESS.2023.3252009.

[57] Z. Wang, S. Yang, H. Qin, Y. Liu, and J. Wang, "MixCFormer: A CNN–Transformer Hybrid with Mixup Augmentation for Enhanced Finger Vein Attack Detection," *Electron.*, vol. 14, no. 2, 2025, doi: 10.3390/electronics14020362.

[58] S. Chaudhari, V. Mithal, G. Polatkan, and R. Ramanath, "An Attentive Survey of Attention Models," *ACM Trans. Intell. Syst. Technol.*, vol. 12, no. 5, pp. 1–33, 2021, doi: 10.1145/3465055.

[59] A. de Santana Correia and E. L. Colombini, *Attention, please! A survey of neural attention models in deep learning*, vol. 55, no. 8. Springer Netherlands, 2022. doi: 10.1007/s10462-022-10148-x.

[60] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[61] S. Albawi, O. Bayat, S. Al-Azawi, and O. N. Ucan, "Social touch gesture recognition using convolutional neural network," *Comput. Intell. Neurosci.*, vol. 2018, 2018, doi: 10.1155/2018/6973103.

[62] Z. C. Lipton, J. Berkowitz, and C. Elkan, "A Critical Review of Recurrent Neural Networks for Sequence Learning," pp. 1–38, 2015, [Online]. Available: http://arxiv.org/abs/1506.00019.

[63] A. Graves, A. R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, no. 3, pp. 6645–6649, 2013, doi: 10.1109/ICASSP.2013.6638947.

[64] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[65] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in Vision: A Survey," *ACM Comput. Surv.*, vol. 54, no. 10, pp. 1–30, 2022, doi: 10.1145/3505244.

[66] S. Artabaz and L. Sliman, "Feature fusion and selection using handcrafted vs. deep learning methods for multimodal hand biometric recognition," *Sci. Rep.*, vol. 15, no. 1, 2025, doi: 10.1038/s41598-025-10075-1.

[67] N. Alay and H. H. Al-Baity, "Deep learning approach for multimodal biometric recognition system based on fusion of iris, face, and finger vein traits," *Sensors (Switzerland)*, vol. 20, no. 19, pp. 1–17, 2020, doi: 10.3390/s20195523.

[68] S. A. El-Rahman and A. S. Alluhaidan, "Enhanced multimodal biometric recognition systems based on deep learning and traditional methods in smart environments," *PLoS One*, vol. 19, no. 2 February, pp. 1–24, 2024, doi: 10.1371/journal.pone.0291084.

[69] J. C. Atenco, J. C. Moreno, and J. M. Ramirez, "Audiovisual Biometric Network with Deep Feature Fusion for Identification and Text Prompted Verification," *Algorithms*, vol. 16, no. 2, 2023, doi: 10.3390/a16020066.

[70] B. H. Kim and J. Y. Pyun, "ECG identification for personal authentication using LSTM-based deep recurrent neural networks," *Sensors (Switzerland)*, vol. 20, no. 11, pp. 1–17, 2020, doi: 10.3390/s20113069.

[71] H. S. Gill, O. I. Khalaf, Y. Alotaibi, S. Alghamdi, and F. Alassery, "Multi-Model CNN-RNN-LSTM Based Fruit Recognition and Classification," *Intell. Autom. Soft Comput.*, vol. 33, no. 1, pp. 637–650, 2022, doi: 10.32604/iasc.2022.022589.

[72] F. U. Cnn and D. N. N. Models, "Multimodal Biometric Recognition Based on Fusion of Electrocardiogram and Multimodal Biometric Recognition Based on Fusion of Electrocardiogram and Fingerprint Using CNN, LSTM, CNN-LSTM, and DNN Models," no. August, 2025, doi: 10.52436/1.jutif.2025.6.4.5098.

[73] S. Ayeswarya and K. J. Singh, "A Comprehensive Review on Secure Biometric-Based Continuous Authentication and User Profiling," *IEEE Access*, vol. 12, no. June, pp. 82996–83021, 2024, doi: 10.1109/ACCESS.2024.3411783.

[74] B. Nithya and P. Sripriya, "Fingerprint Identification by Training a LSTM Network with Fingerprint Segments as Sequence Inputs," *Proc. 6th Int. Conf. Commun. Electron. Syst. ICCES 2021*, pp. 1773–1779, 2021, doi: 10.1109/ICCES51350.2021.9489036.

[75] Y. Huang, H. Ma, and M. Wang, "Multimodal Finger Recognition Based on Asymmetric Networks With Fused Similarity," *IEEE Access*, vol. 11, no. November 2022, pp. 17497–17509, 2023, doi: 10.1109/ACCESS.2023.3242984.

[76] P. Delgado-Santos, R. Tolosana, R. Guest, F. Deravi, and R. Vera-Rodriguez, "Exploring transformers for behavioural biometrics: A case study in gait recognition," *Pattern Recognit.*, vol. 143, p. 109798, 2023.

[77] A. Cosma, A. Catruna, and E. Radoi, "Exploring Self-Supervised Vision Transformers for Gait Recognition in the Wild," *Sensors*, vol. 23, no. 5, 2023, doi: 10.3390/s23052680.

[78] J. N. Mogan, C. P. Lee, K. M. Lim, and K. S. Muthu, "Gait-ViT: Gait Recognition with Vision Transformer," *Sensors*, vol. 22, no. 19, 2022, doi: 10.3390/s22197362.

[79] X. Li, J. Feng, J. Cai, and G. Lin, "FV-MViT: Mobile Vision Transformer for Finger Vein Recognition," *Sensors*, vol. 24, no. 4, 2024, doi: 10.3390/s24041331.

[80] Y. Rui, M. Ling-tao, and Z. Qiu-yu, "AuthFormer: Adaptive Multimodal biometric authentication transformer for middle-aged and elderly people," 2024, [Online]. Available: http://arxiv.org/abs/2411.05395.

[81] Y. Haruna, S. Qin, A. H. Adama Chukkol, A. A. Yusuf, I. Bello, and A. Lawan, "Exploring the synergies of hybrid convolutional neural network and Vision Transformer architectures for computer vision: A survey," *Eng. Appl. Artif. Intell.*, vol. 144, no. January, p. 110057, 2025, doi: 10.1016/j.engappai.2025.110057.

[82] S. Tipper, H. F. Atlam, and H. S. Lallie, "An Investigation into the Utilisation of CNN with LSTM for Video Deepfake Detection," *Appl. Sci.*, vol. 14, no. 21, 2024, doi: 10.3390/app14219754.

[83] J. Priyani, P. Nanglia, P. Singh, V. Shokeen, and A. Sharma, "HGSSA-bi LSTM: A Secure Multimodal Biometric Sensing Using Optimized Bi-Directional Long Short-Term Memory with Self-Attention," *ECS Sensors Plus*, vol. 3, no. 1, 2024, doi: 10.1149/2754-2726/ad1b3a.

[84] S. Cai, Y. Jiang, Y. Xiao, J. Zeng, and G. Zhou, "TransUMobileNet: Integrating multi-channel attention fusion with hybrid CNN-Transformer architecture for medical image segmentation," *Biomed. Signal Process. Control*, vol. 107, no. January, p. 107850, 2025, doi: 10.1016/j.bspc.2025.107850.