Training Alone Together: Developing a Virtual Opponent for Home-Based Kickboxing Training

James-Patrick McCrossan
Independent Researcher, Derby, UK
jpdmccrossan@gmail.com

Katarzyna Stawarz
Cardiff University, Cardiff, UK
StawarzK@cardiff.ac.uk

Solo training leaves martial artists without a resource it cannot replace: a responsive sparring partner - an absence the COVID-19 pandemic made acute when it closed gyms and dojos worldwide. Replicating that partner digitally requires real-time recognition of a practitioner's movement, which is still an open challenge for SportsHCI. Virtual Reality (VR) is one route: consumer headsets can place a digital opponent in front of the practitioner, though few actually adapt to how the user is moving. To address this gap, drawing from our experiences as martial arts practitioners, we present a proof-of-concept VR sparring system for kickboxing that attempts to close a feedback loop of recognition, prediction and response for a solo practitioner. Webcam pose estimation feeds a machine learning (ML) classifier, whose output drives an ML sequence predictor that selects the response of an avatar opponent in Unity, a 3D game engine. The classifier reached 87.22% test accuracy across 18 movement classes but reliably recognised only three in deployment. Assessed through autobiographical design, the system is presented as an early exploration from which we draw two provisional design principles for responsive digital sparring partners: calibrated unpredictability and a readable opponent.

SportsHCI, Virtual Reality, Martial Arts, Pose Estimation, Movement Prediction, Kickboxing

1. INTRODUCTION

A recent synthesis of grand challenges for SportsHCI (Elvitigala et al. 2024) frames the gap this paper addresses: its Sub-Challenge 1.4 identifies how little is known about using real-time sensemaking of bodily performance to support novel digital-physical exercises. It distinguishes two roles for such sensemaking: one after the session, for reflective analysis, and one during it, to adapt the exercise or steer the practitioner's behaviour as they move. A responsive sparring partner is an instance of the second, in-the-moment role, and it is this role we build towards.

Motor learning has long distinguished between 'closed' and 'open' skills (Knapp 1963). Closed skills are performed in a stable, predictable environment under the practitioner's control, like a kata or drilling a jab. Open skills are performed against an unpredictable opponent or environment that must be read continuously, like a wrestling exchange or kickboxing sparring. The two demand related but different cognitive spaces: closed skills reward precision and consistency, while open skills reward anticipation, pattern recognition and decision making under uncertainty (Knapp 1963). In SportsHCI, this distinction surfaces in earlier work on partner-based training. Fogtmann (2007) introduced the term 'kinesthetic empathy interaction' for the design space where bodies read other bodies. In it, one practitioner's bodily anticipation of another is the interaction medium. A skilled fighter often feels the shoulder load before a punch fires; that anticipation is one element of an open skill. Ludvigsen et al. (2010) applied this to elite handball with *TacTowers*, where athletes interact through equipment in real time and interaction depends on contingent reading. In kickboxing, bag work and shadowboxing are essential, but they teach mainly closed-skill components like technique, power, conditioning and combinations. They are less suited to the open skill, because there is little to be empathetic towards.

With gyms and dojos closed during the COVID-19 lockdowns (Bowman 2020), training narrowed to what could be done alone: the closed-skill set. A global survey of martial artists documented widespread disruption to training during the pandemic (Meyer et al. 2021); the authors experienced this disruption first hand, which inspired this work. Although gyms have since reopened, the gap re-emerges whenever a training partner is unavailable, whether between sessions, in remote areas, or on returning to a

© McCrossan et al.

discipline after time away. While solo training could be supported by technology, all options have limitations. Video resources like YouTube tutorials and online courses became a common substitute during the lockdowns (Clark and Lupton 2021) and convey technique. However, the instruction runs in one direction only: the practitioner watches and mimics, with no feedback to correct bad habits as they form. Commercial VR games such as *The Thrill of the Fight*[1] offer an opponent and a workout but target engagement over instruction. The problem is the design, not the medium. VR keeps what sparring depends on. The fighter sees an attack coming in real depth and judges the distance to it, while the opponent stays consistent, controllable and safe to face repeatedly (Bideau et al. 2010). Other fast interceptive sports show this works. Goalkeepers read a virtual thrower better than matched video clips (Vignais et al. 2015), and novices who trained against a virtual opponent on a consumer headset improved at real table tennis (Michalski et al. 2019). *The Thrill of the Fight* does respond to the user, but not in a realistic way. Closing the gap therefore needs an opponent that responds as a trained partner would. In martial arts, these responsive opponents have appeared mainly in karate, most notably Zhang et al.'s *KaraKter* (2018). Comparable systems for other striking arts remain rare, especially on consumer hardware. Building on this, our work pairs real-time technique recognition with a predicted opponent response learned from kickboxing sparring footage.

In this paper, we investigate whether a VR avatar sparring partner can recognise a practitioner's kickboxing technique in real time in order to generate and perform a context-dependent response and attempt to close the sparring loop for a solo practitioner. This work is an exploratory step, with the contribution of an architecture and an account of building it, not a validated training tool.

2. RELATED WORK

2.1. Virtual Reality for Striking Martial Arts

A growing body of work shows VR can support genuine skill development in striking arts, not just exercise. In karate, training in VR alongside conventional practice improves response time (Witte et al. 2022), and response behaviour transfers. Ritter et al. (2022) found response time and response type in kumite statistically similar in VR and reality. A review of VR in karate highlighted benefits for technique, reaction time and decision-making (Lygouras and Tsinakos 2024). What these studies share, however, is a static opponent: the attacks are prerecorded sequences, not reactions to the practitioner. For combat sports specifically, the recurring limitation is the absence of haptic feedback (Li et al. 2025), and the wearables proposed for martial arts training do not provide it, focusing on drill practice instead (Quintero Ovalle et al. 2019).

Research-grade systems have pursued training value since Hämäläinen et al.'s *Kick Ass Kung Fu* (2005) first showed computer-vision-driven martial arts VR was viable, and since then the field has advanced one step at a time, like scoring a learner against a virtual coach (Liu et al. 2020), or adding physical impact by having a robot move a real striking surface into place to meet a user's punch (Mendez et al. 2022). The commercial market is more mature still, led by *The Thrill of the Fight*. It delivers immersion and a real workout, but it does not punish a dropped guard or an over-committed strike the way a trained opponent would.

The closest work to ours is Zhang et al.'s *KaraKter* (2018) where an autonomous VR opponent responds to the practitioner directly. A kumite character tracks the athlete and selects context-appropriate attacks from their behaviour, which high-ranking karate experts judged useful for training. It shows a responsive virtual opponent is achievable, although, like most work in this area, only for karate and on specialised hardware.

2.2. Computer Vision and Machine Learning for Movement Analysis in Combat Sports

Recognising sport movement from sensor or vision input is now a mature paradigm, from early depth-camera systems such as the Microsoft Kinect to today's single-camera pose estimation. This technology often focuses on the areas of classification and prediction.

Cust et al.'s review of 52 studies found supervised classification is the dominant approach across both inertial and vision inputs (Cust et al. 2019). Today, single-camera pose estimation has grown capable enough to recognise striking technique from both offline and live input. The heavier, offline work pushes accuracy, lifting 2D pose estimation towards 3D boxing analysis (Lin et al. 2023) and lighter real time work, like classifying karate techniques from joint angles, can run in real time (Echeverria and Santos 2021), including detecting punches from a single fixed camera (Stefański et al. 2024) or recognising Taekwondo technique with over 95 per cent accuracy (Yang and Wang 2025). However, techniques that differ mainly in how they move over time risk looking alike in any single pose frame. Skeleton-based action recognition treats this as its central problem and models the movement itself. Since ST-GCN first applied graph convolution across

[1] https://store.steampowered.com/app/494150/

whole skeleton sequences (Yan et al. 2018), later models have learned which joints move together rather than assuming it (Shi et al. 2019), attended to frames far apart in time (Plizzari et al. 2021), and stacked pose heatmaps into volumes that tolerate noisy pose estimation (Duan et al. 2022). These sequence models have reached combat sports, where near-identical karate techniques received exactly this treatment (Guo et al. 2021). Prediction has been explored too, though aimed elsewhere, such as Wu and Koike's *FuturePose* (2019) which forecasts a martial artist's own near-future pose for mixed-reality training. Pose estimation can also increasingly stand in for a coach, giving accurate feedback with no expert present (Aulton et al. 2025). Combined, these developments in the field suggest that recognising and predicting striking movement from a single camera in real time is feasible.

Other similar research approaches the gap our work targets without closing it. Mueller et al.'s mediated combat play (2014) preserves body-to-body interaction at a distance but still needs a human partner. Ishac et al.'s Internet-of-Things training mat (2023) runs solo, but gives feedback for closed skill drills rather than a responsive opponent.

In summary, although VR can support sport skill acquisition, autonomous opponents that respond to the practitioner have been shown mainly in karate and on specialised hardware, rather than consumer devices (Section 2.1). Computer vision movement analysis in martial arts has matured across action recognition, pose estimation, action evaluation and athlete support (Pang et al. 2025), yet recognition and prediction have often advanced as separate problems. Our work explores how the two could be joined, and the next section describes the system we built as a step towards this.

3. SYSTEM DESIGN AND DEVELOPMENT

3.1. Positionality: Autobiographical Design

This project is informed by our expertise as martial arts practitioners. The first author has roughly eight years of combined martial arts experience, most of it in kickboxing. The second author is a qualified instructor with a black belt in Choi Kwang Do who also has experience with kickboxing. We have experience of limited training during the pandemic and frustrating attempts to train with VR systems. As a result, through an autobiographical design approach (Neustaedter and Sengers 2012) we served both as designers and expert target users.

3.2. System Design

By design, the system closes a feedback loop with the practitioner through three components: a real-time technique classifier, an ML sequence predictor and a Unity-based VR environment with an animated avatar opponent. A laptop-mounted webcam captures motion. MediaPipe (Lugaresi et al. 2019) converts each frame into 33 body landmarks, from which eight joint angles (left and right elbow, shoulder, hip and knee) are derived. These angles feed the classifier continuously, and classifier outputs accumulate in a rolling buffer of up to three moves. When the avatar finishes its current animation, the predictor consumes this continuously updated buffer and forecasts the avatar's next move. Figure 1 shows the system in deployment and Figure 2 the data flow.

3.2.1. Movement Vocabulary

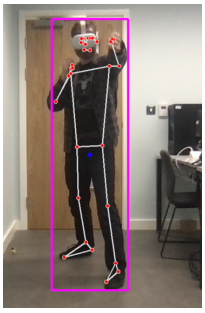
The system targets 18 movement classes: eight punches (jab, cross, hook, uppercut and overhand, the latter three on both sides), four kicks (low and body kicks on each side), five defensive techniques (guard, left and right head blocks, left and right kick blocks) and a neutral stance. The selection reflects techniques that appear frequently in live sparring and compose into the counter-relationships that drive sparring decisions. An incoming jab, for instance, is blocked by a head block, countered by a body kick (if the jabber over-commits), or lands if met only with guard. The vocabulary is illustrative, not exhaustive: head movement defences, clinch work and full footwork variations are out of scope.

3.2.2. Recognition

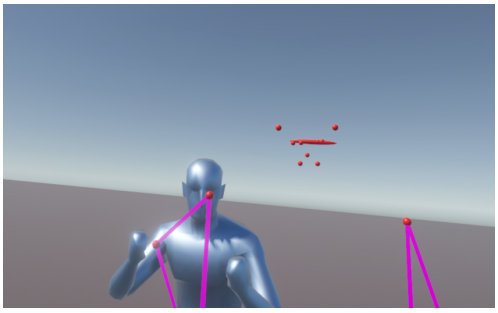
The classifier reads the eight joint angles and returns a probability distribution over the 18 classes. Angles were chosen over raw image pixels for two reasons. First, we expected them to transfer better across different body shapes, in line with skeleton-based action recognition (Echeverria and Santos 2021), though we could not test this with data from a single practitioner. Second, reducing each frame from MediaPipe's 33 landmarks to eight angles kept inference fast enough to run live on a laptop with no GPU. The cost is lost information: weight distribution, torso rotation and the speed of a movement are not captured. The classifier is a hybrid. A K-Nearest Neighbours (K-NN) model makes a first guess, which is added to the eight angles as a ninth input, and a long short-term memory (LSTM) network produces the final label. Training data came from one practitioner performing all 18 movements, giving roughly 11,000 labelled frames; full architecture and training detail are in Appendix A.

3.2.3. Prediction

The predictor takes the three most recent classifier outputs and forecasts the avatar's next move. Like the



***(a)** Input. Webcam feed of the practitioner mid-jab with MediaPipe's 33 landmarks overlaid. The eight derived joint angles feed the classifier.*



***(b)** Output. The first person headset view. The practitioner appears as connected landmark spheres while the avatar opponent faces them at sparring distance and plays the response the predictor selects.*

***Figure 1:** System in deployment: the webcam input with pose overlay (left) and the first person view the practitioner sees in the headset (right).*

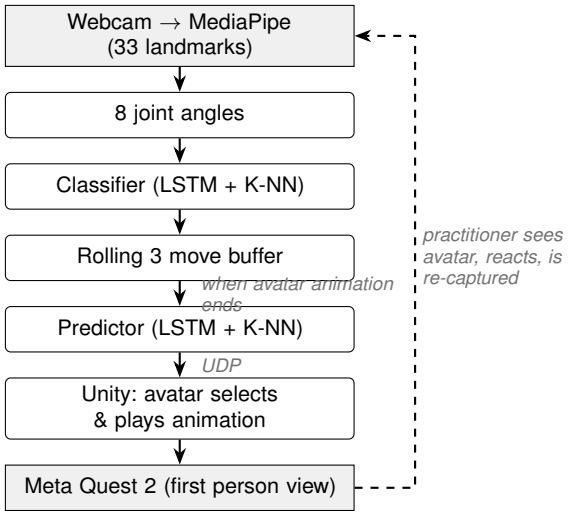


***Figure 2:** System architecture: the recognition-prediction-response loop between practitioner and avatar. The solid path is the per-action data flow; the dashed path closes the loop as the practitioner reacts to the avatar and is re-captured by the webcam.*

classifier, it pairs a K-NN step with an LSTM. It runs only when the avatar finishes its current animation, rather than on a fixed clock. Its training data came from 168 fight clips drawn from the kickboxing category of the YouTube-8M dataset (Abu-El-Haija et al. 2016). Both fighters' poses were estimated by MoveNet (TensorFlow Team 2021). Crucially, the classifier itself labelled every frame of this footage, so the data the predictor learned from carried the classifier's own judgements. Its architecture and training settings are given in Appendix A.

3.2.4. VR Implementation

The Unity scene presents the practitioner with a first person view of the avatar at sparring distance and a real-time visualisation of their own body (see Figure 1b). That body is drawn as 33 connected spheres at the MediaPipe landmark coordinates, with line segments between adjacent joints forming a moving skeleton.[2] Spheres were simpler and less computationally demanding than a rigged avatar mesh, which needs calibration and skinning. They still gave a clear reference for the limbs in space, legs included. A fully rigged avatar driven by the same landmarks is a natural next step.

Avatar locomotion used distance based behaviour: the opponent retreats when the practitioner is too close, advances when too far, strafes laterally at mid-range and triggers an attack animation within striking range once a per-action cooldown has elapsed. The attack played is whichever the predictor returned. Distance thresholds, strafe speed and cooldown duration are exposed as Unity inspector fields and adjustable without code changes. Animations were drawn from commercial Unity Asset Store packs.[3]

3.2.5. Deployment and Latency

The whole pipeline runs on a single laptop: a 2018 MacBook Pro with an Intel i7, 16 GB of RAM and no dedicated GPU. It streams to the Meta Quest 2 over an iPhone 11 hotspot. Each inference cycle takes upwards of 200 ms, and every practitioner action needs two of them: classifying the current pose, then calling the predictor. The avatar therefore responds roughly 400 ms after the practitioner moves, enough to noticeably degrade the sparring loop. That is right at the 400 ms tipping point Durnez et al. (2022) identify for exergames, where performance and experience begin to degrade sharply. For finer motor VR tasks such as aiming, the tolerances are tighter: accuracy degrades significantly from roughly 90 ms when target motion is unpredictable, or roughly 130 ms when it can be anticipated (Kelkkanen et al. 2023). The cost looks like a deployment problem rather than a design one. Inference runs on a laptop

[2] Embodiment approach inspired by "Real-Time Motion Capture with MediaPipe and Unity" (Murtaza's Workshop, 2022).

[3] Fighting Motions Vol. 1 (Magicpot Inc., 2016); Combat Animations - Kickboxing and Muay Thai v1 and V2 (Ahmed Saad, 2023); obtained via the Unity Asset Store, `https://assetstore.unity.com`.

with no GPU, and the result travels over a phone hotspot. The approach requires two inferences per action, but each would be faster on better hardware.

3.3. Expert Assessment

Without a controlled user study at this stage, we assess the system at two levels: a quantitative evaluation of the machine learning components, and a structured first-author assessment grounded in autobiographical design (Neustaedter and Sengers 2012), judged against four requirements a sparring partner must meet: embodiment in the virtual space, an opponent that responds to the practitioner, responses that are valid kickboxing, and real-time interaction. The first author set the system up, performed kickboxing techniques, and reacted to the avatar's responses. The recognition loop closed end-to-end and the avatar's responses were valid kickboxing moves, but the system fell short as a sparring partner: embodiment through the wireframe body was thin, the move vocabulary in deployment was narrow, and a response delay of roughly 400 ms (Section 3.2.5) let the practitioner out-pace the avatar's strikes. The deeper limitation is that the avatar could be reacted to but not read. It never feinted, its face never changed, and it never loaded a shoulder or shifted weight before striking. Attacks began at full speed from a neutral stance, so there was nothing to anticipate. Its replies also became more predictable within the collapsed movement vocabulary, predictable enough to bait rather than spar against. Without these cues, the reading of an opponent that defines an open skill had nothing to act on.

3.3.1. What Building the System Surfaced

When the practitioner throws a technique, the classifier recognises it from their body angles. On test data it reached 87.22% accuracy across all 18 classes, but in deployment its reliable output collapsed to three (jab, cross and left body kick). The cause was the representation: judging eight joint angles one frame at a time kept inference fast but separated only the most distinct standing poses, so techniques that differ mainly in how they move over time were lost. In effect, the classifier seems to recognise poses rather than movements: a ceiling set by the representation, which more training data or tuning alone seems unlikely to lift. This is the failure mode temporal sequence models address (Section 2).

3.3.2. Reading the Predictor's Behaviour

Once the classifier has recognised the practitioner's move, it feeds the predictor, which chooses how the avatar replies. The predictor reached roughly 60% accuracy, but that accuracy percentage needs careful interpretation. First, the predictor did not learn from footage labelled by a human expert. As described in Section 3.2.3, its training data were fight clips labelled by the classifier itself. The classifier could only reliably recognise three moves, so those three labels dominated the training data. The predictor therefore learned to predict within that narrow set rather than the full 18: a weakness in the first stage became a limitation in the second stage. We report this cascade as an observed pattern rather than a quantified effect. A full decomposition is beyond the scope of a short paper. A low accuracy score here is not in itself a fault, however: a fully predictable opponent can be exploited, so some unpredictability is wanted rather than avoided. We expand on this point in Section 4.

4. DISCUSSION

We asked whether a solo practitioner could spar against a virtual opponent that recognises their kickboxing moves in real time. Informed by our expertise as martial arts practitioners through autobiographical design, we built a prototype that reads movement from a single webcam, classifies the technique, predicts a response and plays it back through an avatar in VR. We discuss what the system suggests for designing responsive sparring partners, then the limitations and future work it points to.

4.1. Designing Responsive Sparring Partners

The assessment points to two provisional design principles specific to responsive sparring partners. The first concerns the level of challenge. As the assessment showed, once the avatar's replies became predictable they could be baited rather than trained against, so a sparring partner's value does not lie in predicting the next move as accurately as possible. One that is too erratic is too hard, with no pattern to learn from. Effective training plausibly sits between the two, the middle ground that research on game difficulty and immersion finds most engaging (Sweetser and Wyeth 2005). This reframes prediction as something to calibrate rather than maximise, and distinguishes a learned predictor from a system built for a single discipline such as *KaraKter* (Zhang et al. 2018): unpredictability becomes a property to shape in the model. Variability of this kind is a long-established prerequisite for transferable open-skill learning under the contextual interference framework (Shea and Morgan 1979), so the target is task-relevant variability rather than maximal predictiveness.

The second principle surfaced in use: the opponent must be readable. A fighter anticipates by reading another body, the load of a shoulder before a punch, a shift of weight, a flicker in the face, exactly the

bodily reading that Fogtmann's (2007) kinesthetic empathy interaction depends on. As our system was intended as a proof-of-concept at this stage, the avatar produced movement without any of these cues, so it gave a trainee plenty to react to but nothing to anticipate, leaving out the layer of open-skill training that depends on reading an opponent. A responsive sparring partner should therefore model not only what the opponent does but the bodily tells that let a trainee see it coming. Reading these tells depends on depth and a first person view, which a flat screen cannot give and VR can (Vignais et al. 2015).

4.2. Limitations and Future Work

The most significant limitation is that the system has not been studied with practitioners other than the authors. While autobiographical design is a valid HCI method (Neustaedter and Sengers 2012), every claim about how an anticipatory virtual opponent might support sparring is, at this stage, a hypothesis. Future work should evaluate this approach with other users. Nevertheless, we could still identify opportunities and limitations of this system, which provide valuable insights to other researchers working in this area.

Second, the classifier's training data came from a single practitioner, which bounds generalisation across body types, stances and styles, although this is in line with similar exploratory research (Quintero Ovalle et al. 2019). Nevertheless, this is a clear limitation. Broader, more varied training data would be expected to help the classifier generalise across body types and styles, especially if anticipatory facial and movement cues are to be added.

Third, the effective vocabulary in deployment was three techniques (jab, cross and left body kick) rather than 18 - enough to demonstrate the architecture, not enough to constitute real sparring. Addressing this likely means utilising a temporal sequence model (Section 2), the standard fix for this ceiling, and confirming the new model stays reliable as the move set grows. Such a model costs more per inference and so sharpens the latency constraint (Section 3.2.5). That constraint remains a deployment limit, and lighter pose models, better hardware or a less heavily chained inference path can keep the system real time.

A full per-class evaluation, with precision, recall and a confusion matrix, is future work, as is widening the input beyond vision-only joint angles through multimodal fusion with inertial data (Yadav et al. 2022). Then, with a reliable, wider vocabulary in place, the predictor should be validated against annotated expert sparring data, so its forecasts can be judged. Only after the completion of the above does the final step become impactful: a comprehensive study with kickboxers, to test if training against an anticipatory VR opponent develops the reactive, open skills that bag and pad work leave untouched.

5. CONCLUSION

We have presented a proof-of-concept VR sparring system for kickboxing, built from our own expertise as practitioners. It combines real-time pose estimation, a hybrid LSTM/K-NN classifier and an ML sequence predictor to drive an avatar that reacts to the practitioner, closing the recognition-prediction-response loop end to end on consumer hardware. Our contribution is an early account of bringing ML recognition and prediction into a single interactive loop for a digital-physical exercise, in a narrow vocabulary rather than a finished sparring partner, as a first step towards one. From building and assessing it, we draw two provisional design principles for responsive sparring partners: the need to calibrate unpredictability rather than maximise accuracy, and to make the opponent readable so a trainee can anticipate its movements. Whether training against such an anticipatory opponent develops the open skills that solo practice cannot is the question this work opens, and the one a fuller version will need to answer in studies with kickboxers.

REFERENCES

Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. YouTube-8M: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016.

Cavan Aulton, Lois Wakili, Ben William Strafford, Keith Davids, and Chuang-Yuan Chiu. The application of deep learning human pose estimation in sport: A systematic review. *Sports Medicine - Open*, 11(1): 155, 2025. doi: 10.1186/s40798-025-00953-3.

Benoit Bideau, Richard Kulpa, Nicolas Vignais, Sébastien Brault, Franck Multon, and Cathy Craig. Using virtual reality to analyze sports performance. *IEEE Computer Graphics and Applications*, 30(2): 14–21, 2010. doi: 10.1109/MCG.2009.134.

Paul Bowman. Metamorphoses of martial arts: Meditations on motivations and motive forces during the COVID-19 pandemic. *Martial Arts Studies*, (10):128–140, 2020. doi: 10.18573/mas.119.

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002. doi: 10.1613/jair.953.

Marianne Clark and Deborah Lupton. Pandemic fitness assemblages: The sociomaterialities and affective dimensions of exercising at home during the COVID-19 crisis. *Convergence*, 27(5):1222–1237, 2021. doi: 10.1177/13548565211042460.

Emily E. Cust, Alice J. Sweeting, Kevin Ball, and Sam Robertson. Machine and deep learning for sport-specific movement recognition: A systematic review of model development and performance. *Journal of Sports Sciences*, 37(5):568–600, 2019. doi: 10.1080/02640414.2018.1521769.

Haodong Duan, Yue Zhao, Kai Chen, Dahua Lin, and Bo Dai. Revisiting skeleton-based action recognition. In *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2969–2978, 2022. doi: 10.1109/CVPR52688.2022.00298.

Wouter Durnez, Aleksandra Zheleva, Mark Claypool, Mathias Maes, Klaas Bombeke, Jan Van Looy, and Lieven De Marez. Spaz! the effects of local latency on player actions in a desktop-based exergame. *IEEE Transactions on Games*, 14(4): 623–631, 2022. doi: 10.1109/TG.2021.3128714.

Jon Echeverria and Olga C. Santos. Toward modeling psychomotor performance in karate combats using computer vision pose estimation. *Sensors*, 21(24): 8378, 2021. doi: 10.3390/s21248378.

Don Samitha Elvitigala, Armağan Karahanoğlu, Andrii Matviienko, Laia Turmo Vidal, Dees Postma, Michael Jones, Maria F. Montoya, Daniel Harrison, Lars Elbæk, Florian Daiber, Lisa Anneke Burr, Rakesh Patibanda, Paolo Buono, Perttu Hämäläinen, Robby van Delden, Regina Bernhaupt, Xipei Ren, Vincent van Rheden, Fabio Zambetta, Elise van den Hoven, Carine Lallemand, Dennis Reidsma, and Florian 'Floyd' Mueller. Grand challenges in SportsHCI. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24. Association for Computing Machinery, 2024. doi: 10.1145/3613904.3642050.

Maiken Hillerup Fogtmann. Kinesthetic empathy interaction – exploring the possibilities of psychomotor abilities in interaction design. In *Proceedings of the Second International Workshop on Physicality (Physicality 2007)*, pages 37–42. UWIC Press, 2007.

Jianping Guo, Hong Liu, Xi Li, Dahong Xu, and Yihan Zhang. An attention enhanced spatial–temporal graph convolutional LSTM network for action recognition in karate. *Applied Sciences*, 11 (18):8641, 2021. doi: 10.3390/app11188641.

Perttu Hämäläinen, Tommi Ilmonen, Johanna Höysniemi, Mikko Lindholm, and Ari Nykänen. Martial arts in artificial reality. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '05, pages 781–790. Association for Computing Machinery, 2005. doi: 10.1145/1054972.1055081.

Karlos Ishac, Katia Bourahmoune, and Marc Carmichael. An IoT sensing platform and serious game for remote martial arts training. *Sensors*, 23 (17):7565, 2023. doi: 10.3390/s23177565.

Viktor Kelkkanen, David Lindero, Markus Fiedler, and Hans-Jürgen Zepernick. Hand-controller latency and aiming accuracy in 6-DOF VR. *Advances in Human-Computer Interaction*, 2023:1563506, 2023. doi: 10.1155/2023/1563506.

Barbara Knapp. *Skill in Sport: The Attainment of Proficiency*. Routledge & Kegan Paul, 1963.

Yike Li, Hansen Li, Chun Jiang, Yuqin Su, Sijia Jiang, and Guodong Zhang. Advancements in virtual reality for performance enhancement in combat sports: A mini-review and perspective. *Frontiers in Psychology*, 16:1563212, 2025. doi: 10.3389/fpsyg.2025.1563212.

Jianchu Lin, Xiaolong Xie, Wangping Wu, Shengpeng Xu, Chunyan Liu, Toshboev Hudoyberdi, and Xiaobing Chen. Model transfer from 2D to 3D study for boxing pose estimation. *Frontiers in Neurorobotics*, 17:1148545, 2023. doi: 10.3389/fnbot.2023.1148545.

Juan Liu, Yawen Zheng, Ke Wang, Yulong Bian, Wei Gai, and Dingyuan Gao. A real-time interactive Tai Chi learning system based on VR and motion capture technology. *Procedia Computer Science*, 174:712–719, 2020. doi: 10.1016/j.procs.2020.06.147.

Martin Ludvigsen, Maiken Hillerup Fogtmann, and Kaj Grønbæk. TacTowers: an interactive training equipment for elite athletes. In *Proceedings of the 8th ACM Conference on Designing Interactive Systems*, DIS '10, pages 412–415. Association for Computing Machinery, 2010. doi: 10.1145/1858171.1858250.

Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. MediaPipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019. URL `https://arxiv.org/abs/1906.08172`.

Dimosthenis Lygouras and Avgoustos Tsinakos. The use of immersive technologies in karate training: A scoping review. *Multimodal Technologies and Interaction*, 8(4):27, 2024. doi: 10.3390/mti8040027.

Luis Andres Mendez, Ho Yin Ng, Zin Yin Lim, Yi-Jie Lu, and Ping-Hsuan Han. MovableBag: Substitutional robot for enhancing immersive boxing training with encountered-type haptic. In *SIGGRAPH Asia 2022 XR*. Association for Computing Machinery, 2022. doi: 10.1145/3550472.3558406.

Martin J. Meyer, Andrea Molle, Benjamin N. Judkins, and Paul Bowman. Martial arts in the pandemic. *Martial Arts Studies*, (11):7–31, 2021. doi: 10.18573/mas.134.

Stefan Carlo Michalski, Ancret Szpak, Dimitrios Saredakis, Tyler James Ross, Mark Billinghurst, and Tobias Loetscher. Getting your game on: Using virtual reality to improve real table tennis skills. *PLoS ONE*, 14(9):e0222351, 2019. doi: 10.1371/journal.pone.0222351.

Florian 'Floyd' Mueller, Martin Gibbs, Frank Vetere, Stefan Agamanolis, and Darren Edge. Designing mediated combat play. In *Proceedings of the 8th International Conference on Tangible, Embedded and Embodied Interaction*, TEI '14, pages 149–156. Association for Computing Machinery, 2014. doi: 10.1145/2540930.2540937.

Carman Neustaedter and Phoebe Sengers. Autobiographical design in hci research: designing and learning through use-it-yourself. In *Proceedings of the Designing Interactive Systems Conference*, DIS '12, pages 514–523. Association for Computing Machinery, 2012. doi: 10.1145/2317956.2318034.

Yiqun Pang, Yibing Wang, Qiurui Wang, Fengmei Li, Changnian Zhang, and Chuanwei Ding. Applications of AI in martial arts: A survey. *Proceedings of the Institution of Mechanical Engineers, Part P: Journal of Sports Engineering and Technology*, 239(2):301–331, 2025. doi: 10.1177/17543371241273827.

Chiara Plizzari, Marco Cannici, and Matteo Matteucci. Skeleton-based action recognition via spatial and temporal transformer networks. *Computer Vision and Image Understanding*, 208–209:103219, 2021. doi: 10.1016/j.cviu.2021.103219.

Juan Quintero Ovalle, Katarzyna Stawarz, and Asier Marzo. Exploring the addition of audio input to wearable punch recognition. In *Proceedings of the XX International Conference on Human Computer Interaction*, Interacción '19, pages 1–4. Association for Computing Machinery, 2019. doi: 10.1145/3335595.3335641.

Yvonne Ritter, Melina Droste, Dan Bürger, Stefan Pastel, and Kerstin Witte. Comparison of response behavior in karate kumite between real world and virtual reality. *Sports Engineering*, 25(1):14, 2022. doi: 10.1007/s12283-022-00378-1.

John B. Shea and Robyn L. Morgan. Contextual interference effects on the acquisition, retention, and transfer of a motor skill. *Journal of Experimental Psychology: Human Learning and Memory*, 5(2):179–187, 1979. doi: 10.1037/0278-7393.5.2.179.

Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12018–12027, 2019. doi: 10.1109/CVPR.2019.01230.

Piotr Stefański, Jan Kozak, and Tomasz Jach. Boxing punch detection with single static camera. *Entropy*, 26(8):617, 2024. doi: 10.3390/e26080617.

Penelope Sweetser and Peta Wyeth. GameFlow: A model for evaluating player enjoyment in games. *Computers in Entertainment*, 3(3):3, 2005.

TensorFlow Team. MoveNet: Ultra fast and accurate pose detection model. Google TensorFlow Hub, 2021. URL `https://www.tensorflow.org/hub/tutorials/movenet`.

Nicolas Vignais, Richard Kulpa, Sébastien Brault, Damien Presse, and Benoit Bideau. Which technology to investigate visual perception in sport: Video vs. virtual reality. *Human Movement Science*, 39:12–26, 2015. doi: 10.1016/j.humov.2014.10.006.

Kerstin Witte, Melina Droste, Yvonne Ritter, Peter Emmermacher, Steffen Masik, Dan Bürger, and Katharina Petri. Sports training in virtual reality to improve response behavior in karate kumite with transfer to real world. *Frontiers in Virtual Reality*, 3:903021, 2022. doi: 10.3389/frvir.2022.903021.

Erwin Wu and Hideki Koike. FuturePose: Mixed reality martial arts training using real-time 3D human pose forecasting with a RGB camera. In *Proceedings of the 2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1384–1392, 2019. doi: 10.1109/WACV.2019.00152.

Santosh Kumar Yadav, Aditya Deshmukh, Raghurama Varma Gonela, Shreyas Bhat Kera, Kamlesh Tiwari, Hari Mohan Pandey, and Shaik Ali Akbar. MS-KARD: A benchmark for multimodal karate action recognition. In *Proceedings of the 2022 International Joint Conference on Neural Networks*

(IJCNN), pages 1–8. IEEE, 2022. doi: 10.1109/IJCNN55064.2022.9892646.

Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, AAAI '18, pages 7444–7452. AAAI Press, 2018. doi: 10.1609/aaai.v32i1.12328.

Feng Yang and Zizhuo Wang. An intelligent taekwondo coaching system based on augmented reality technology with real-time feedback mechanisms. *Scientific Reports*, 15(1):40832, 2025. doi: 10.1038/s41598-025-24608-1.

Liang Zhang, Guido Brunnett, Katharina Petri, Marco Danneberg, Steffen Masik, Nicole Bandow, and Kerstin Witte. KaraKter: An autonomously interacting karate kumite character for VR-based training and research. *Computers & Graphics*, 72: 59–69, 2018. doi: 10.1016/j.cag.2018.01.008.

STATEMENT ON AI USE

AI (Anthropic's Claude) was used to aid writing by assisting in refining human drafted text and reducing word counts, but it was not used in any other part of this project.

A. MODEL ARCHITECTURES AND TRAINING DETAILS

Data collection. The classifier's training data were recorded over two sessions in the same week, with rest between movements. Each of the 18 movements was performed for 5-7.5 minutes per session, 10-15 minutes and roughly 30-50 repetitions in total across the two sessions, while the practitioner rotated through 360 degrees to vary the camera's perspective.

Classifier. A five neighbour K-Nearest Neighbours model produces a first pass label, which is appended to the eight joint angles as a ninth value. This nine-value per-frame vector is reshaped into a nine length sequence and passed to a stacked LSTM. The LSTM has a bidirectional 64-unit layer, then 128- and 64-unit layers, followed by dense layers of 128, 64 and 32 units and a softmax over the 18 classes. The LSTM therefore runs over the feature axis. Training used a 70/30 train/test split with SMOTE (synthetic minority over-sampling technique) for class balance (Chawla et al. 2002), the Adam optimiser and categorical cross entropy loss, to a maximum of 100 epochs with an early stopping patience of 10. Training stopped at epoch 58, restoring the best weights from epoch 48. On the held-out test set the model reached 87.22 per cent accuracy (loss 0.7694); training accuracy at the stopping point was 92.99 per cent (loss 0.4370).

Predictor. The predictor operates on move labels, not joint angles. Its input is the sequence of the practitioner's three most recent classified moves; a five neighbour K-Nearest Neighbours prediction is appended to that window, giving a four-step input to a stacked LSTM. The LSTM has a bidirectional 64-unit layer, then 128- and 64-unit layers, with dropout and L2-regularised dense layers, ending in a softmax over the move set. The training sequences were built from 168 YouTube-8M clips: MoveNet (TensorFlow Team 2021) estimated both fighters' poses and the classifier labelled every frame, so the predictor learned from the classifier's own move labels rather than from poses. Training used an 80/20 split and sparse categorical cross-entropy loss, with class balance handled by weighted resampling rather than the classifier's SMOTE, to a maximum of 100 epochs with an early stopping patience of 10. Training stopped at epoch 27, restoring the best weights from epoch 17. On the held out test set the restored model reached 60.29 per cent accuracy (loss 1.0989); training accuracy at the stopping point was 49.02 per cent (loss 1.0441).