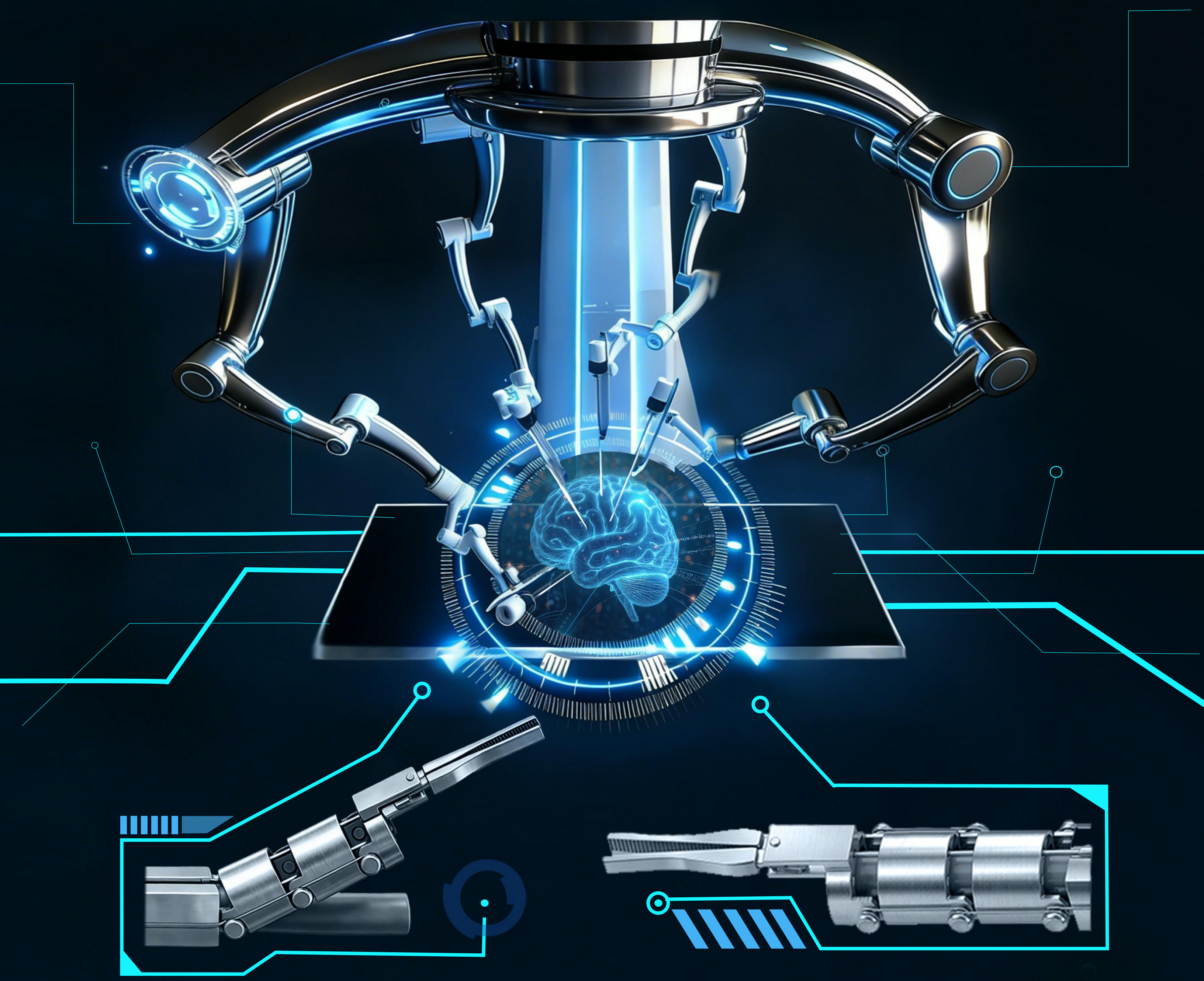
Open Access

PROGRESS IN MEDICAL DEVICES

ISSN: 2957-5478

Editor-in-Chief: Haipo Cui



Volume 4 , Issue 1 · 2026

Published by ZENTIME PUBLISHING CORPORATION

Progress in
Medical Devices

RESEARCH ARTICLE

Development of an automated cytological smear staining device for rapid on-site evaluation

Guangyan Wang[1], Kai Yang[1], Chunhua Zhou[2], Duowu Zou[2], Shiju Yan[1]

[1]School of Health Science and Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China.
[2]Department of Gastroenterology, Ruijin Hospital, Shanghai 200025, China.

Corresponding author: Shiju Yan.

Address correspondence to: Shiju Yan, School of Health Science and Engineering, University of Shanghai for Science and Technology, No. 516 Jungong Road, Yangpu, Shanghai 200093, China. E-mail: yanshiju@usst.edu.cn.

Received January 20, 2025; Accepted April 16, 2025; Published March 24, 2026

DOI: 10.61189/599339cpncph

Abstract

Objective: To develop an automated staining device addressing the issues of cumbersome operation, low efficiency, and cluttered workspace associated with current clinical rapid on-site evaluation using Diff-Quik staining for aspiration specimen cytological smears. **Methods:** The device integrates a microcontroller to control motors, peristaltic pumps, solenoid valves, and fans, enabling the automatic transfer of cytological slides, delivery of staining and rinsing solutions, and forced-air drying. **Results:** Preliminary test results indicated that the developed device successfully performs four key steps, including Diff-Quik A and B staining, rinsing, and drying, achieving automatic staining of cytological smears. **Conclusion:** The developed device features a compact footprint and has no pollution to the operating environment, offering a practical solution for automated, efficient, and convenient slide staining during rapid on-site evaluation procedures.

Keywords: Rapid on-site evaluation, Cytological smear, Staining, Automation, Device development

Highlights

- The developed device reduces manpower and time consumption, improving staining efficiency in digestive endoscopy centers.
- It has a compact design with minimal contamination to the operating environment.
- The developed device demonstrates excellent staining performance and has been recognized by clinicians.

1 INTRODUCTION

In 2022, pancreatic cancer accounted for approximately 511,000 new cases and 467,000 deaths worldwide. It remains one of the most lethal malignancies, ranking sixth in cancer-related mortality, contributing to nearly 5% of all cancer deaths worldwide. The mortality rate of pancreatic cancer has remained stable in many countries, including those in the European Union, over the last decades. As mortality from other cancers, such as lung cancer, colorectal cancer, prostate cancer, breast cancer and gastric cancer, continues to decline, the public health burden of pancreatic cancer has increased [1, 2]. Over one-third of pancreatic cancer cases from 1990 to 2019 were associated with the digestive system [3].

For the diagnosis of pancreatic and other digestive tract lesions, endoscopic ultrasound-guided fine-needle aspiration/biopsy (EUS-FNA/B) has become a widely accepted and effective technique [4-6]. EUS-FNA/B enables real-time, ultrasound-guided needle puncture to obtain specimens from target lesions, offering advantages of minimal trauma, rapid recovery, and low cost [7].

During EUS-FNA/B procedures, Rapid On-Site Evaluation (ROSE) facilitates pathologists to quickly assess the adequacy and diagnostic value of collected specimens, which is particularly crucial for less experienced operators and centers with lower diagnostic yields. In current clinical practices, ROSE is

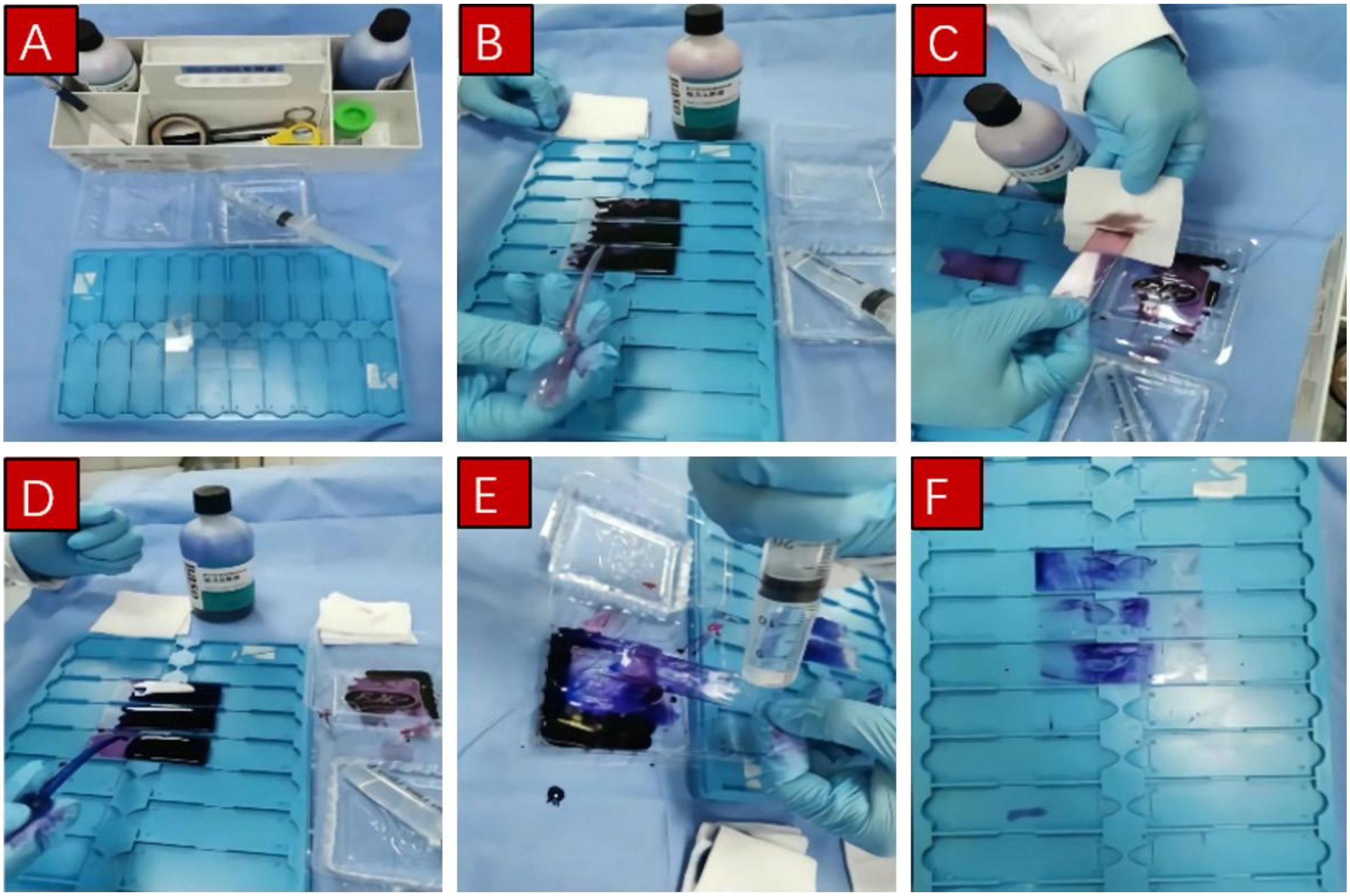


Figure 1. Flow chart of Diff-Quick staining of cytological smears.

typically performed using Diff-Quik staining, a rapid staining method evolved from Wright's stain and is commonly applied in cytological evaluations. Diff-Quik staining offers a quick turnaround time and produces clear background staining, aiding in accurate visualization of cellular morphology and structure [8, 9]. Through Diff-Quik staining, pathologists can preliminarily determine whether aspirated specimens is representative of the lesion and evaluate its pathological properties [10].

Currently, the Diff-Quik staining process is manually operated by medical personnel (**Figure 1**). Manual operations, including smear preparation, fixation, staining, and rinsing, have several drawbacks: 1. Low efficiency, long time consumption, and significant physical workload; 2. Space occupation by equipment and tools, along with environmental disruption caused by manual dripping of dye solution and rinsing under running water; and 3. High demand for experience and skills.

These challenges highlight an urgent clinical need for an automated cytological smear processing device to enhance staining efficiency, reduce environmental contamination, and enhance the practicality of Rose in routine workflows.

In this study, we developed an automated cytological smear staining device for ROSE. This device is capable of independently executing four core processes - Diff-Quik A and B staining, rinsing, and air drying, and is equipped with safety features such as emergency stop and power-off restart to ensure reliable operation and minimize potential losses in clinical settings.

2 MATERIALS AND METHOD

2.1 Overall scheme design

The developed cytological smear staining device is based on the STM32G432RBT6 microcontroller and integrates multiple functional units, such as LCD display, control buttons, stepper motors, a peristaltic pump, and solenoid valves. The schematic diagram of the device composition is shown in **Figure 2**.

The device supports sequential staining, real-time process visualization, power-off data retention, and emergency braking. Its operation is initiated by the user via control buttons, which trigger the microcontroller to actuate the stepper motor,

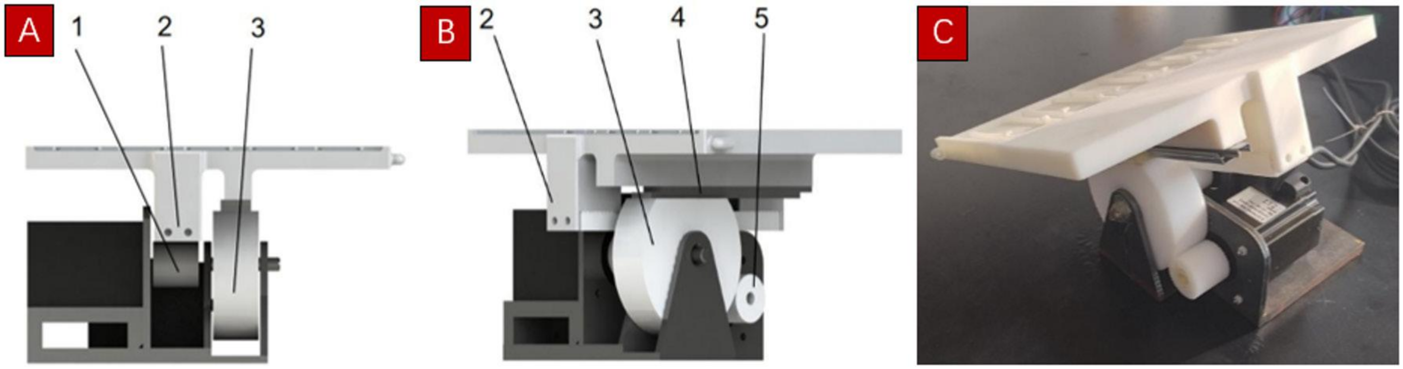


Figure 2. Structure of the smear carrier and its base. (A) Main view of the structure; (B) Side view of the structure; (C) Physical prototype of the structure. 1: gear 1, 2: rack fixedly connected below the carrier, 3: gear 3, 4: guideway, 5: gear.

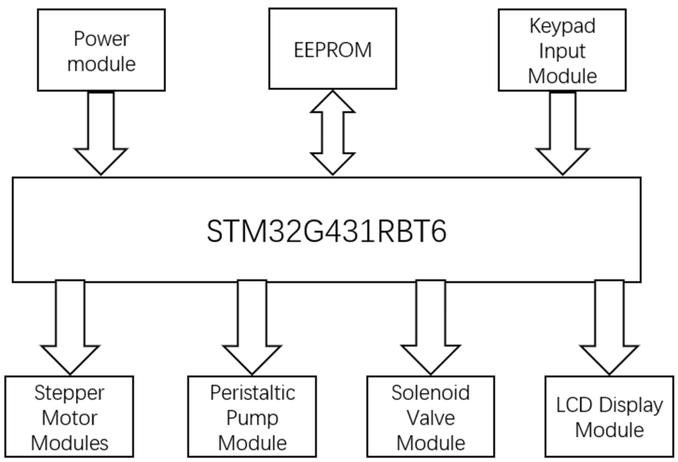


Figure 3. Diagram of the overall system components.

enabling reciprocating motion of the carrier. Meanwhile, the peristaltic pumps transport dye solution, with solenoid valves controlling switch between different dye solution. Real-time staining status is displayed in on the LCD screen. During the staining process, position data of both the motor and the carrier are stored in Electrically Erasable Programmable Read Only Memory (EEPROM), allowing process continuity in the event of a power outage. The waiting time between steps is programmable, enabling optimization of staining quality and precise control of the smear drying duration.

2.2 Mechanical structure design

2.2.1 Design of carrier motion mechanism

The mechanical structure of the device was designed using SolidWorks. As shown in **Figure 3**, the carrier motion mechanism is essentially a two-degree-of-freedom variable pivot rotation system [11]. In this design, the carrier supports cytological smear slides and performs two types of motion: translation and rotation. The translation is achieved using a gear-rack mechanism, which converts the rotational motion of a motor into linear movement of the carrier. The rotation is achieved through a gear mechanism, which converts the rotational motion of the motor into the tilting motion of the carrier. The motion transmission routes are as follows:

Translation: The output shaft of the translation motor is fixed to gear 1, which engages with rack 2. Rack 2, rigidly attached beneath the carrier, moves linearly, driving the carrier along the horizontal axis. During linear motion, gear 3 remains stationary, and its flattened upper surface remains in contact with the bottom of the moving carrier, allowing relative motion between the carrier and gear 3.

Rotation: The output shaft of the rotation motor is fixed to gear 5, which meshes with gear 3. When gear 3 rotates, it induces a tilting motion in the carrier. During this process, the bottom of the carrier maintains contact with the flattened upper part of gear 3, but no relative motion occurs between them.

To avoid mechanical interference between translation and rotation, gear 1 and gear 3 are designed to be coaxial.

The base of the carrier is made of steel to ensure sufficient load-bearing capacity. Two stepper motors and all associated gears are mounted on this base. The output shaft of each motors is connected to its relevant driving gear by a key. After receiving the motion command from the upper computer, the microcontroller activates the translation motor, driving the carrier to undergo horizontal linear movement. Once the carrier reaches the specified position, the rotation motor is activated to tilt the carrier. Simultaneously, the peristaltic pump operates during the translation phase to deliver staining or rinsing solution. During the tilting phase, the peristaltic pump stops working to drain residual staining or rinsing solution, preventing cross-contamination and ensuring staining quality.

2.2.2 Design of the liquid pathway

Three different liquid containers are used to store dye solution A, dye solution B, and rinsing water, respectively. The outlet of

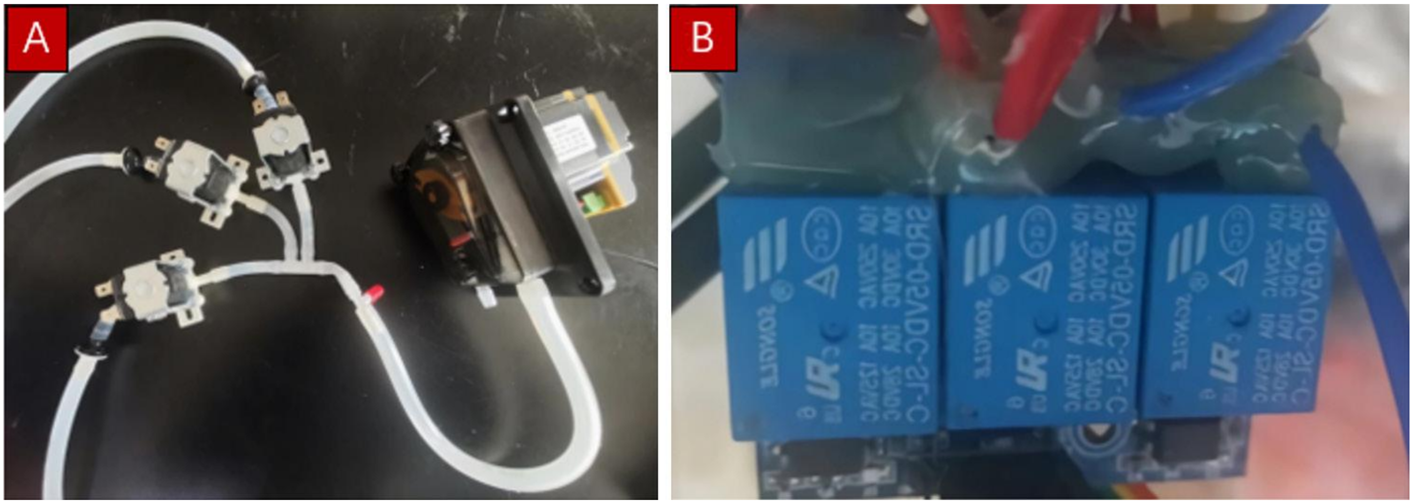


Figure 4. The liquid pathway and the relay. (A) Peristaltic pump and solenoid valves connected to the liquid pathway; (B) Three-way relay.

each container is connected to a four-way pipe through its corresponding solenoid valves. These individual lines are merged into a single output pipe. All three types of liquid are driven by a peristaltic pump and finally drained to the cytological smears fixed on the carrier through a dropper. Solenoid valves are individually controlled via a relay module. **Figure 4A** shows the configuration of the peristaltic pump and solenoid valves within the liquid pathway system, while **Figure 4B** shows the relay module for value control.

2.3 Component selection and software design

2.3.1 Component selection

In this study, Altium Designer was used to design the circuit schematic and Printed Circuit Board layout. The electronic control system of the developed automated smear staining device consists of several functional modules, including the main control module, stepper motor module, peristaltic pump module, solenoid valve control module, LCD display module, EEPROM storage module, button input module, and power supply module.

In the main control module, a STM32G431 chip is used for overall system control and data processing. This microcontroller offers a stable development environment and an extensive set of peripheral modules, enabling seamless integration with external hardware components.

In the stepper motor module, a type 57 stepper motor paired with a DM556 stepper motor driver was selected [12-14]. The stepper motor provides a holding torque of 2.4 N·m, which is sufficient to support stable movement of the carrier.

For the peristaltic pump module, a 304 K peristaltic pump driven by a JBT-BM controller was used [15, 16]. The pump operates at a rotation speed of 300 r/min and is compatible with #24 tubing. The flow rate of the pump is set at 1000 ml/min, which is sufficient for the delivery of dye solution and rinsing water.

For the solenoid valve module, a normally closed, direct-acting solenoid valve operating at DC 24 V was selected [17, 18]. The pressure range of this solenoid valve is 0-0.2 MPa. Its opening and closing are controlled via a three-way relay module.

In this device, a TFT-LCD screen is incorporated to display real-time information related to staining process and stepper motor positioning, especially for emergency detection and failure handling [19]. This LCD screen operates at 5 V and is capable of displaying 10 lines of information simultaneously, with a maximum of 16 characters per line. The LCD screen features a backlight with a working voltage of 4.5 to 5.5 V, ensuring clear display even under low-light conditions.

To prevent parameter loss and reset difficulties in the event of a power outage, the system stores the stepper motor position in real time. A 2K-bit EEPROM (AT24C02) was selected for this purpose [20]. The EEPROM operates independently of program space and achieves data reading and writing through Inter-Integrated Circuit communication technology, ensuring safe data storage in case of a power outage.

The button module enables user interaction and manual intervention during the staining process. Each button is connected to GPIO port on the microcontroller, ensuring efficient hardware wiring and convenient software processing.

Specifically, buttons B1, B2, B3, and B4 are connected to pins PB0, PB1, PB2, and PA0 of the microcontroller, respectively. B1 initiates the staining processes, B2 confirms smear placement, B3 confirms smear removal, and B4 has dual functions - short press for emergency stop and long press for execution

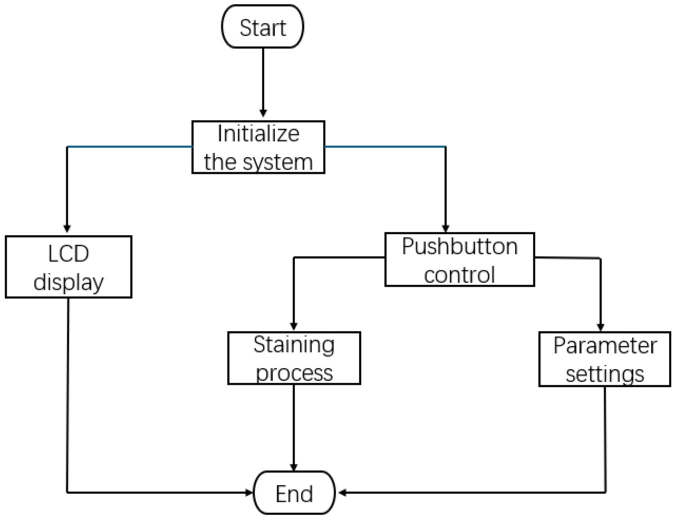


Figure 5. Process flow chart.

mode switching. Once the system is powered on, the microcontroller continuously scans the button states. When a button is pressed, the corresponding program is executed promptly.

2.3.2 Software design

The software program was developed using the Keil uVision5 platform from Keil Corporation. Software control is the key to stable and coordinated operation of all functional modules. The control system includes subprograms, such as the main program, stepper motor driver, peristaltic pump controller, storage program, button interaction program, and solenoid valve controller. As shown in **Figure 5**, the program first performs initialization of all modules after startup. Then, the operating mode of the device is selected and controlled via physical button inputs. During the staining process, the LCD display module continuously displays the current operational stage in real time.

The button control program plays a central role in realizing in realizing human-machine interaction, enabling manual setting and adjustment of the device's operational state.

The function update_Key_Status () is responsible for real-time button status updates. It is embedded in the system tick handler SysTick_Handler (), ensuring continuous scanning of button states after the system is powered on. Once a button press is detected, the corresponding task program is executed immediately.

The output I/O ports of the microcontroller are connected to the control signal terminals of the DM556 driver: pulse (PUL)+, PUL-, direction (DIR)+, DIR-, enable+, and enable-. The PUL and DIR signals are used to control the motor's motion and direction, while the enable signals control motor activation. In order to prevent interference caused by the simultaneous operation of two motors, the translation motor is given higher control priority than the rotation motor during execution.

The control program for peristaltic pumps is relatively simple. By configuring the GPIO ports responsible for the pump's power and direction, the microcontroller can initiate or terminate the fluid delivery process as required.

Three relays are used to control the corresponding solenoid valves. To prevent malfunctions caused by negative pressure during valve operation, an LED status indicator is incorporated to display real-time operational condition of the peristaltic pump.

In the storage program, Inter-Integrated Circuit bus is used for data communication. The Inter-Integrated Circuit bus protocol includes five basic data transmission modes. In this device, Current Address Read mode is applied for reading, and Byte Write mode is used for data writing [21].

The LCD display program is designed to provide real-time feedback on device status. In parameter setting mode, the LCD displays the configured operating speed of the stepper motor. In dyeing mode, it displays the current dyeing step, the remaining time for that step, and relevant user operation information.

Upon system startup, all modules undergo initialization. Subsequently, the smear staining process is initiated, and the LCD displays operation instructions simultaneously. In standby mode, the buttons are continuously monitored, and the staining process will be activated after the user responds. The stepper motor drives the platform to execute translation and rotation, the peristaltic pump to supply liquid, and the relay-controlled solenoid valves to switch the dye solution.

The flow of the staining process is shown in **Figure 6**.

Upon startup, all modules are initialized and enters a standby state, waiting for subsequent commands (button B1);

1. Once button B1 is pressed, the staining process starts. The microcontroller controls the stepper motor to move the carrier to the front access window, allowing users to open the device and load cytological smears onto the carrier. The control system then waits for the staining initiation command (button B2);

2. After the smears are loaded and button B2 is pressed, the carrier is moved to the rear position directly below the dropper. The peristaltic pump and the solenoid valve for staining solution A are activated;

3. The dropper begins dispensing staining solution A, while the carrier moves forward simultaneously to ensure even coverage of the smear;

4. Upon completion of the first staining step, the carrier returns to the rear position and tilts to drain residual liquid;

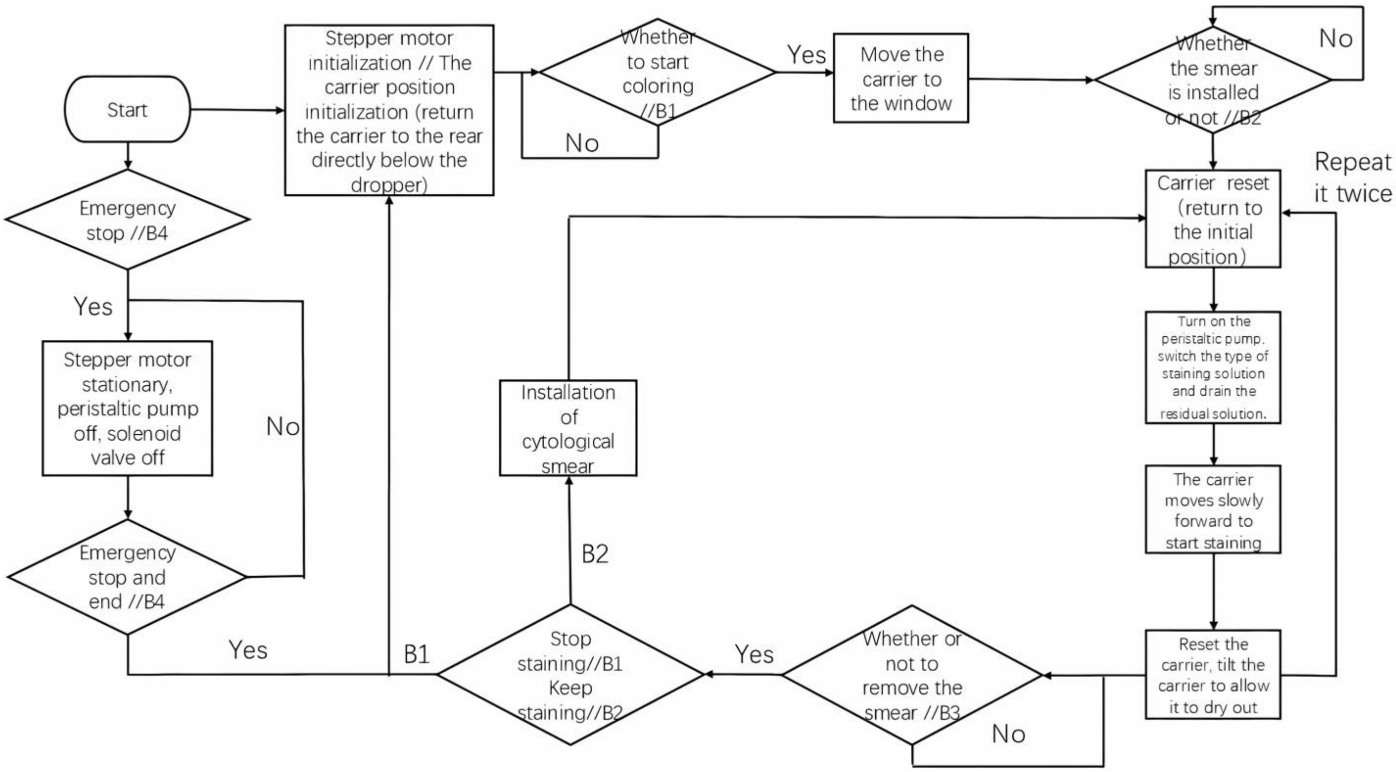


Figure 6. Staining process flow chart.

Figure 7. Physical prototype of the staining device.

5. The air-drying process is initiated. Simultaneously, the solenoid valve for staining solution A is closed, and the solenoid valve for staining solution B is opened; The above steps are then repeated for solution B and subsequent rinsing or drying processes as required;

6. After all staining and drying steps are completed, the system awaits the smear removal command (button B3);

7. When button B3 is pressed, the carrier is moved to the front window position, allowing users to open the device and retrieve the stained smears.

If a device malfunction occurs during any stage, button B4 may be pressed to enter sleep mode. In this mode, all system components, solenoid valves, peristaltic pump, and stepper motors, are shut off to prevent further error or damage. Once the fault is repaired, button B4 can be pressed again to restart the initialization procedure and allow the staining process to resume from the beginning.

2.4 Verification of device functions

2.4.1 Validation scheme

Initially, the hardware circuit underwent rigorous inspection to ensure precise soldering and reliable connections. Subsequently, all modules were integrated into the complete device assembly. As shown in **Figure 7**, the staining device is housed within a rectangular acrylic enclosure, the bottom and side panels are

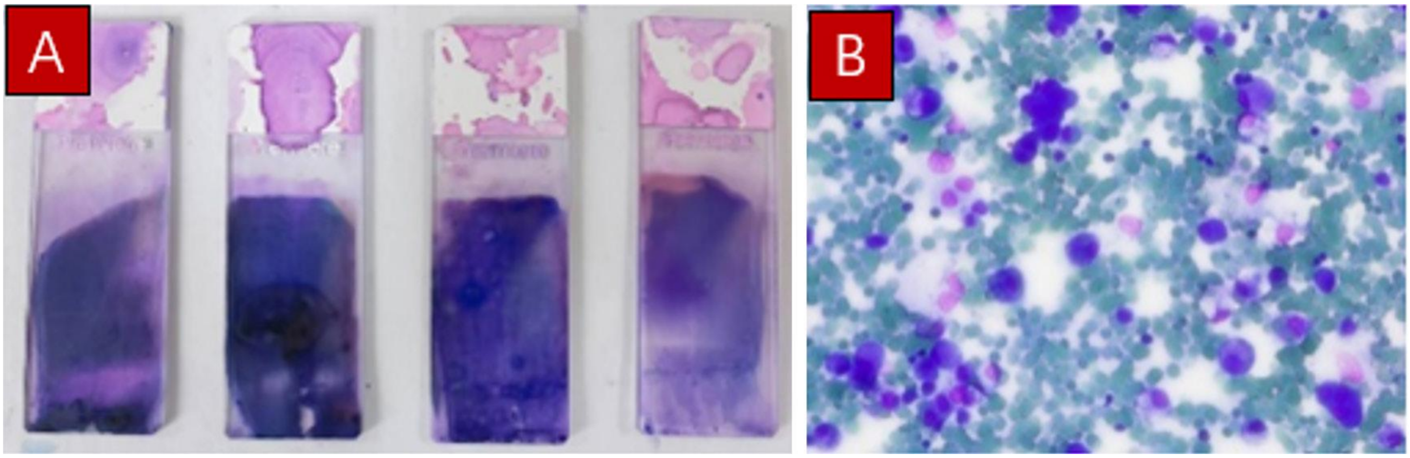


Figure 8. Visual and microscopic examination results of stained cytological smear. (A) Four stained cytological smears; (B) Microscopic image of a cytological smears.

sealed to prevent external interference or dye leakage, and a partition plate is installed inside the housing, separating the compartments: the main mechanical structure is centrally mounted on the bottom, while the electronic control hardware is installed on the partition plate above. The liquid pipeline and droppers are placed directly above the carrier. The dye and rinsing solution reservoirs are located at the rear inside the housing.

Following assembly, individual module tests were performed to verify each component's function. After confirming the proper operation of all modules, comprehensive system-level testing was carried out to validate its performance and expected functions.

2.4.2 Test of the modules

The button module was tested by observing the response of an LED indicator on the microcontroller. The LED control code was embedded at the points corresponding to short and long button presses to verify accurate detection. When the relevant GPIO port was pulled low (grounded), the LED turned on. To prevent mutual interference between the LED module and the LCD module, a 74HC573 latch was used for electrical isolation. During LED control, the latch enable bit was first set high to enable the latch, then pulled low to maintain the LED's state.

The LED control snippet is as follows:

```
HAL_GPIO_WritePin(LEDEN_GPIO_Port,LEDEN_Pin,
GPIO_PIN_SET);
HAL_GPIO_WritePin(LED0_GPIO_Port,LED0_Pin,GPIO_
PIN_RESET);
HAL_GPIO_WritePin(LEDEN_GPIO_Port,LEDEN_Pin,
GPIO_PIN_RESET);
```

After successful verification of the button module functionality, testing proceeded to the stepper motor module, peristaltic pump, and solenoid valve module using the button inputs. Each button was pressed individually to confirm that the corresponding device functions operated correctly and without error.

During operation of the stepper motor, press the RESET button on the microcontroller at any time and observe the variation of the display parameter jiaodu_a on the LCD screen before and after pressing the RESET button. If the value of jiaodu_a remained unchanged before and after pressing, it confirmed that the EEPROM module correctly stored the position data and retained it after a reset, thus passing the test.

Next, the staining process test was carried out. The staining process code was executed, and key operations were performed according to the prompts on the LCD display module. The variation of information displayed on the LCD was observed at each step of the staining process, ensuring that the system provided real-time feedback to the user.

The testing results show that the microcontroller effectively controlled the stepper motors, the peristaltic pump, and the solenoid valves to stain the smear on the carrier with different staining solutions.

3 RESULTS

After successful testing of all modules, an overall staining test was performed on the staining device. A and B Diff staining, rinsing, and drying of several hydrothorax cytological smears were carried out sequentially. The hydrothorax cytological smears specimens were provided by the Digestive Endoscopy Department of Shanghai Ruijin Hospital. **Figure 8** shows the visual and microscopic examination results of the stained smear. The ROSE operating pathologist at Shanghai Ruijin

Hospital evaluated the smear staining quality and confirmed that the quality of staining was sufficient for microscopic examination and diagnosis.

4 DISCUSSION AND CONCLUSIONS

In response to the current challenges of low efficiency, cluttered operating spaces, and high manpower consumption in Diff-Quick staining of cytological smears, we designed an automatic staining device specifically for ROSE-oriented aspiration specimen cytological smears. The prototype device underwent comprehensive functional verification, demonstrating its ability to effectively automate the four key processes: A and B Diff staining, rinsing, and drying of smears. The developed staining device offers several advantages, including enhanced efficiency, improved workspace organization, emergency stop functionality, and process monitoring.

In summary, this cytological smear staining device is a valuable tool for rapid staining and on-site evaluation of cytological smears, particularly in fine needle aspiration guided by endoscopic ultrasound, such as those used in pancreatic lesion diagnosis. It holds promise for improving the implementation of EUS-FNA/B cell rapid staining technology in digestive endoscopy centers.

Looking ahead, large-scale clinical trials will be conducted in collaboration with multiple medical institutions to further refine device parameters. Future efforts will focus on validating the device's applicability to other sample types, such as lymph node punctures and thyroid fine-needle aspirations. Additionally, integration of additional staining techniques, such as Papanicolaou staining and immunocytochemical staining, will be explored to address while expanding diagnostic capabilities through.

DECLARATIONS

Author contributions

Guangyan Wang: Collect and summarize materials, write and revise the paper, and draw graphs and tables; Kai Yang: Consult literature, and revise the content of the paper; Chunhua Zhou: Conduct critical review of the informative content of the article; Duowu Zou: Explain experimental data and evaluate experimental results; Shiju Yan: Design related experiments, Provide writing ideas and guide paper writing.

Funding

This research received no external funding.

Data availability

Not applicable.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

Not applicable.

REFERENCES

[1] Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2024 May-Jun;74(3):229-263. https://doi.org/10.3322/caac.21834

[2] Zhang X, Yang L, Liu S, Cao L, Wang N, Li H, et al. Interpretation on the report of global cancer statistics 2022. Chin J Oncol. 2024 Jul 23;46(7):710-721. https://doi.org/10.3760/cma.j.cn112152-20240416-00152

[3] Wang Y, Huang Y, Chase RC, Li T, Ramai D, Li S, et al. Global burden of digestive diseases: A systematic analysis of the global burden of diseases study, 1990 to 2019. Gastroenterology. 2023 Sep;165(3):773-783.e715. https://doi.org/10.1053/j.gastro.2023.05.050

[4] Tian L, Tang AL, Zhang L, Liu XW, Li JB, Wang F, et al. Evaluation of 22G fine-needle aspiration (FNA) versus fine-needle biopsy (FNB) for endoscopic ultrasound-guided sampling of pancreatic lesions: a prospective comparison study. Surg Endosc. 2018 Aug;32(8):3533-3539. https://doi.org/10.1007/s00464-018-6075-6

[5] Wang BG, Lee IH. EUS-FNA diagnosis of pancreatic tophaceous gout: Two rare cases. Diagn Cytopathol. 2024 Sep; 52(9):E222-E225. https://doi.org/10.1002/dc.25323

[6] Majumdar K, Sakhuja P. Surgical Pathology of the gastrointestinal system: Volume I - gastrointestinal tract. 1th ed. Singapore: Springer Singapore; 2022. chapter 6, Diagnostic special stains, immunohistochemical markers, and special techniques used in gastrointestinal tract pathology; p. 113-150.

[7] Wu M, Zhang J. Research progress on auxiliary techniques related to endoscopic ultrasound-guided fine needle aspiration/biopsy. J Nantong Univ (Med Ed). 2022;42(02):137-141. https://doi.org/10.16424/j.cnki.cn32-1807/r.2022.02.010

[8] Jimenez-Heffernan JA, Alvarez F, Muñoz-Hernández P, Bárcena C, Azorin D, Bernal I, et al. Cytologic features of ventricular tumors of the central nervous system: A review with emphasis on diff-quik stained smears. Acta Cytol. 2021;65(2):111-122. https://doi.org/10.1159/000512723

[9] Silverman JF, Frable WJ. The use of the diff-quik stain in the immediate interpretation of fine-needle aspiration biopsies. Diagn Cytopathol. 1990;6(5):366-369. https://doi.org/10.1002/dc.2840060516

[10] Iswenti N, Aric Frendi A, Nicen S. Application of leadership style to patient safety culture in hospitals. Proceedings of 1st

International Conference on Health Sciences and Biotechnology (ICHB 2021); 2022. Atlantis Press. p. 86-88.

[11] Zhang M, Wang Y, Li S, Tao J, Xue H, Chai D, et al. Research on rotation control method of X-ray gastrointestinal diagnosis bed based on single chip microcomputer. Software. 2019;40(04):100-102. https://doi.org/10.3969/j.issn.1003-6970.2019.04.021

[12] Liu B, Cheng S. Stepper motor and drive control system. Harbin: Harbin Institute of Technology Press; 1997.

[13] Cheng D. Mechanical Design Handbook. Beijing: Electronic Industry Press; 2016.

[14] Fan C, Fan W. Selection and calculation of stepper motor. Mach Tools Hydraul. 2008;(05):310-313,324.

[15] Gao H, Liu T, Sun Z. The Peristaltic pump's principle and application in chemical mechanical polishing process. Equip Electron Prod Manuf. 2010;39(09):48-51.

[16] Ferretti P, Pagliari C, Montalti A, Liverani A. Design and development of a peristaltic pump for constant flow applications. Front Mech Eng. 2023 July 18;Volume 9 - 2023. https://doi.org/10.3389/fmech.2023.1207464

[17] Dong Y. Analysis and use of solenoid valves. J Salt Sci Chem Ind. 2020;49(03):45-46,52. https://doi.org/10.16570/j.cnki.issn1673-6850.2020.03.014

[18] Wang X. Principle and selection of solenoid valves. Mach China. 2014;(11):183.

[19] Zhang H, Li H, Zhong H. TFT-LCD liquid crystal display technology and application research. Electron Software Eng. 2022;(18):86-89. https://doi.org/10.20109/j.cnki.etse.2022.18.021

[20] Xiang S. Study of key circuits for reliable and low power EEPROM memory in RFID [master' thesis]. Chengdu: Southwest Jiaotong University; 2018.

[21] Xu Z, Li W, Hua Y, Xu C. Application of I^2C bus technology in I/O port expansion of ship energy management system. Chin Ship Res. 2019;14(01):144-149. https://doi.org/10.19693/j.issn.1673-3185.01129

Progress in
Medical Devices

REVIEW ARTICLE

Review of key technologies in ankle rehabilitation robots

Jiajia Zha, Qingyun Meng, Hongtao Shen, Mingxia Wei

School of Health Science and Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China.

Corresponding author: Qingyun Meng.

Address correspondence to: Qingyun Meng, School of Health Science and Engineering, University of Shanghai for Science and Technology, No. 516 Jungong Road, Yangpu District, Shanghai 200093, China. Tel: +86-13761813609. E-mail: mengqy@sumhs.edu.cn.

Received May 24, 2025; Accepted July 25, 2025; Published March 24, 2026

DOI: 10.61189/730741lcujht

Abstract

Ankle rehabilitation robots represent an important branch of rehabilitation robotics, offering significant potential to improve the quality of life for patients with ankle dysfunction caused by stroke, sports injuries, and other conditions. This review first outlines the anatomy and range of motion of the ankle joint, compares conventional rehabilitation approaches with robot-assisted therapy, and highlights the clinical significance of ankle rehabilitation robots. It then systematically examines current research progress from two core perspectives: mechanical structure design and control strategies. In mechanical design, the performance characteristics of series versus parallel mechanisms are compared, the advantages and limitations of actuation methods such as electric motors and pneumatic artificial muscles are analyzed, and the application contexts of platform-based and wearable robots are discussed. In control strategies, the discussion covers motion control and human-robot interaction, beginning with fundamental position, velocity, and trajectory tracking control, and extending to intention-level and cognitive interaction. Finally, based on current research and clinical needs, future ankle rehabilitation robots are expected to evolve toward greater flexibility, intelligence, and universality, providing a theoretical foundation for future studies.

Keywords: Ankle rehabilitation robot, Mechanical design, Control strategy, Human-Robot interaction

Highlights

- As a primary weight-bearing joint, the ankle is highly susceptible to injury, while neurological disorders such as stroke can further impair its motor function, leading to long-term gait disturbances.
- Rehabilitation robots can be platform-based or wearable: platforms aid early-stage motion restoration, while wearable designs focus on gait retraining.
- Control systems must prioritize motion accuracy and safety. Adaptive algorithms boost performance, while bioelectric signal integration enables intention recognition. Coupling with virtual or augmented reality further enhances patient engagement.

1 INTRODUCTION

As the joint with the highest load-bearing capacity in the human body, the ankle has a complex physiological structure. Its bony anatomy and degrees of freedom (DOF) are shown in **Figure 1** [1]. The ankle joint is primarily composed of the tibia, fibula, talus, and calcaneus, and has three DOF: dorsiflexion/plantarflexion (DO/PL) in the sagittal plane, adduction/abduction (AD/AB) in the horizontal plane, and inversion/eversion (IN/EV) in the coronal plane.

The ankle is highly susceptible to both acute injuries and chronic disorders. As the primary load-bearing joint of the skeleton, it sustains forces approximately five times body weight during normal walking and up to 13 times body weight during running [2]. Consequently, ankle sprains are among the

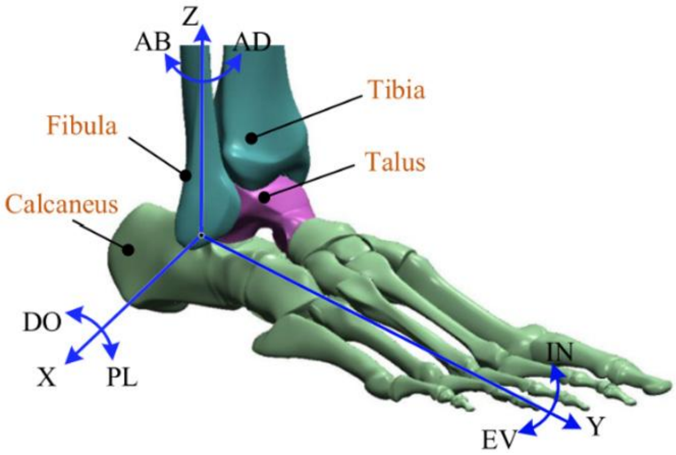


Figure 1. Bony structure and degrees of freedom (DOF) of the ankle joint. This figure is cited from [1]. AD/AB, Adduction/Abduction; IN/EV, Inversion/Eversion; DO/PL, Dorsiflexion/Plantarflexion.

most common sports injuries, accounting for about 7-10% of all cases [3]. Anatomically, the medial malleolus is formed by the distal tibia, whereas the lateral malleolus is formed by the distal fibula. Because the lateral malleolus is positioned lower and the medial deltoid ligament is stronger than the lateral ligament complex, the ankle is more prone to inversion injuries [4].

Notably, up to 50% of patients with acute ankle sprains do not seek formal medical care, leading to recurrent sprains over subsequent months or years. Even with structured rehabilitation, up to 70% experience persistent functional impairment within 6 months, progressing to functional ankle instability [5].

In addition to trauma, chronic neurological conditions—particularly stroke—can severely impair ankle motor function [6]. According to the World Health Organization, the proportion of the global population aged ≥60 years will rise from 12% in 2015 to 22% in 2050, reaching 2.1 billion people [7]. As stroke prevalence increases with age, the incidence of stroke-related ankle dysfunction is expected to rise accordingly.

Ankle injuries, which often result in long-term pain and functional limitations even after the initial recovery, restrict daily activities and diminish quality of life [8]. Patients often face gait instability, balance deficits, and difficulty walking, substantially reducing independence. Therefore, restoring motor function and mobility is a clinical priority.

Rehabilitation training has been shown to improve post-stroke outcomes, prevent muscle atrophy and deep vein thrombosis, and enhance mobility [9, 10]. Early rehabilitation after ankle injury is essential to minimize the adverse effects of prolonged bed rest. Current rehabilitation programs typically include stretching, gait training, constraint-induced movement therapy, and other neuroplasticity-enhancing interventions [11]. However, conventional rehabilitation requires intensive involvement from medical teams and family members, and is limited by constrained healthcare resources, high costs, low efficiency, and inconsistent movement quality [12].

Robot-assisted therapy offers a promising solution to these challenges [13]. It enables more effective, standardized rehabilitation under medical supervision, mitigates workforce shortages, reduces costs, and improves training efficiency. Furthermore, rehabilitation robots can collect detailed kinematic and kinetic data to monitor progress. When combined with virtual reality (VR), such systems can enhance patient engagement and motivation, benefiting both physical and psychological recovery [14-16].

This review summarizes advances in ankle rehabilitation robot design and control strategies based on a comprehensive literature survey. Section 2 discusses mechanical design classifications, Section 3 examines control strategy classifications, and Section 4 presents conclusions and future perspectives.

2 MECHANICAL STRUCTURE DESIGN

The mechanical structure of an ankle rehabilitation robot is a key determinant of its motion performance and therapeutic efficacy. Based on design approach, current systems can be classified as series, parallel, or hybrid structures. Series configurations are lightweight and mechanically simple but have limited capacity to replicate the ankle's complex motions. Parallel configurations offer high precision and large torque output but have a restricted range of motion. Hybrid configurations combine the strengths of both, balancing flexibility and stability, though their structural complexity often leads to classification under either predominantly series or parallel designs.

The choice of actuation method—such as electric motors or pneumatic muscle actuators (PMAs)—also influences flexibility and energy efficiency. Motor drives are more suitable for late-stage rehabilitation requiring high precision and rapid response, whereas PMAs are preferred in early-stage rehabilitation where safety and compliance are prioritized.

In early rehabilitation, rigid platform-based devices can provide stabilizing support and improve ankle range of motion through passive and active training modes. During gait recovery, flexible wearable robots facilitate real-time, dynamic assistance to restore a normal gait cycle. **Table 1** summarizes the structural characteristics, suitable rehabilitation stages, advantages, and limitations of platform-based versus wearable ankle rehabilitation robots.

2.1 Structural form

2.1.1 Series structure

Lee et al. developed a compact cable-driven series elastic actuator (SEA) capable of lightweight design and high-precision torque output, achieved through gait-phase–synchronized

Table 1. Comparison between platform ankle rehabilitation robot and wearable rehabilitation ankle robot

Ways to use	Platform ankle rehabilitation robot	Wearable ankle rehabilitation robot
Structural features	Rigid structure, often using multi-degrees of freedom parallel mechanism, with certain support force	Flexible and lightweight exoskeleton structure
Drive mode	Mainly motor driven, less pneumatic artificial muscles	Mainly pneumatic artificial muscles, less motor drive
Applicable rehabilitation stage	Suitable for early passive training and mid-term active resistance training, etc.	Suitable for mid- to late-stage gait training and daily activity assistance, etc.
Advantages	Multi-degrees of freedom coupling training capability is strong, output accuracy is high, and load capacity is large	Highly portable, comfortable to wear, supports real-time dynamic gait assistance
Limitation	The system is large and expensive, and is not suitable for home rehabilitation scenarios	Low load capacity, limited freedom, short battery life

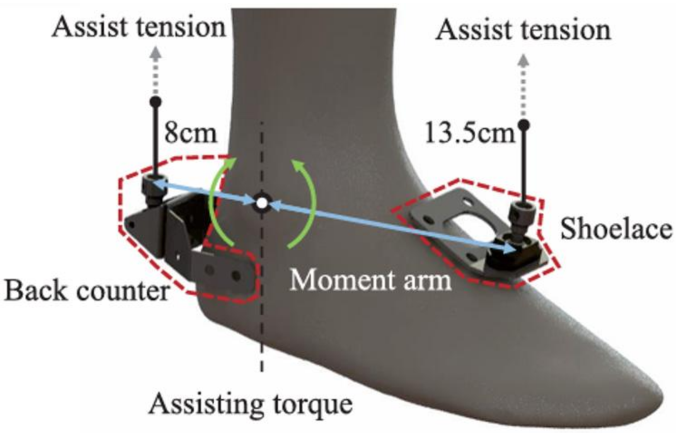


Figure 2. Compact cable driven series flex modules. This figure is cited from [17].

Figure 3. Personalized ankle exoskeleton based on series elastic actuator. This figure is cited from [18]. AFO, Ankle Foot Orthosis.

feedforward friction compensation and nonlinear spring control [17]. This device targets foot drop and propulsion deficits in stroke patients but does not provide full ankle joint DOF (**Figure 2**).

Eraky et al. designed a personalized ankle exoskeleton employing an SEA with cable drive and 3D scanning–based customization, enabling both lightweight construction and high torque output [18]. However, it lacks full ankle DOF coverage and requires individualized parameter adjustments (**Figure 3**).

Choi et al. proposed an SEA driven by a flexible shaft, achieving low inertia and precise torque tracking through nonlinear stiffness linearization and disturbance observer control [19]. Nevertheless, it has limited bandwidth and does not encompass the full DOF of the ankle joint (**Figure 4**).

2.1.2 Parallel structure

Wang et al. designed a novel 2-UPS/RRR parallel ankle rehabilitation robot (PARR) and developed a height-adjustable mobile platform to align with the ankle rotation center of different patients (**Figure 5**) [20]. The 2-UPS/RRR mechanism consists of two UPS active branches and one RRR passive branch, forming a 3 DOF parallel mechanism. Each UPS branch comprises a universal (U) joint, a prismatic (P) joint, and a spherical (S) joint in series, while the RRR branch consists of three revolute (R) joints in series.

Huo et al. developed a full-cycle PARR with large angular range, high torque, and multi-DOF coupled rehabilitation capabilities [21]. The design achieves complete constraint through rigid branches without applying additional force to the ankle, employs a cable drive, and offers three DOF (**Figure 6**).

Xie et al. proposed a 3 DOF dual-ankle rehabilitation robot based on a 3-SPS/RUR parallel mechanism with adjustable inter-ankle distance [22]. The system supports both circular and linear rehabilitation trajectories (**Figure 7**). Here, 3-SPS/RUR denotes a parallel mechanism comprising three SPS active branches and one RU*R passive branch, where the underline indicates a passive kinematic pair.

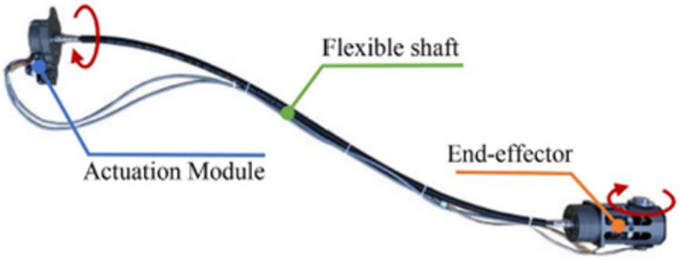


Figure 4. Flexible series elastic actuator. This figure is cited from [19].

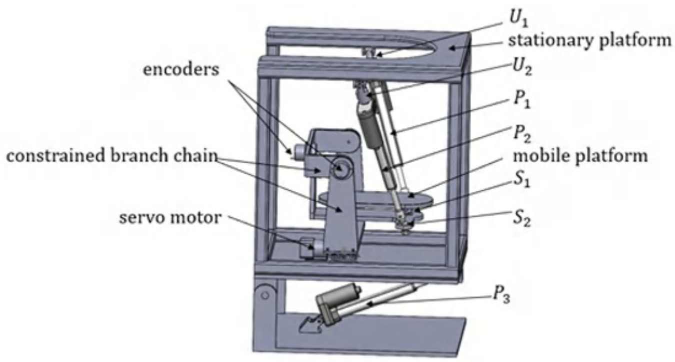


Figure 5. 2-UPS/RRR parallel ankle rehabilitation robot. This figure is cited from [20]. U, Universal; P, Prismatic; S, Spherical; R, Revolute.

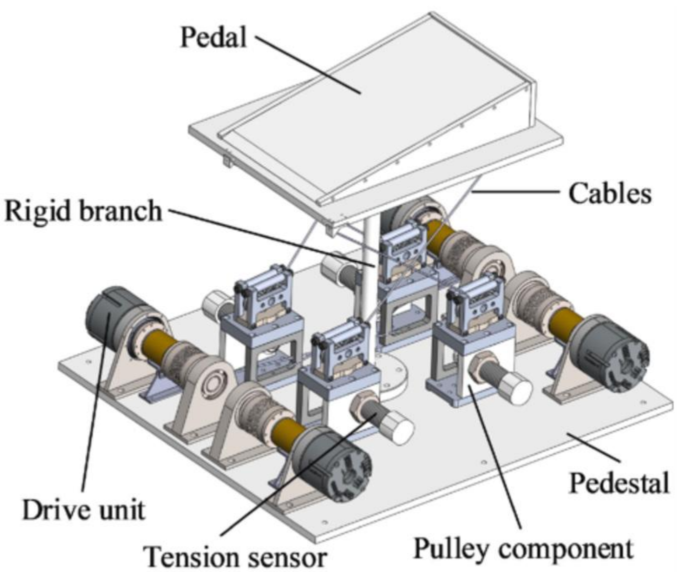


Figure 6. Full-cycle parallel ankle rehabilitation robot. This figure is cited from [21].

2.2 Drive mode

2.2.1 Motor drive

Motors convert electrical energy into mechanical energy via an electromagnetic field, delivering torque or rotational speed.

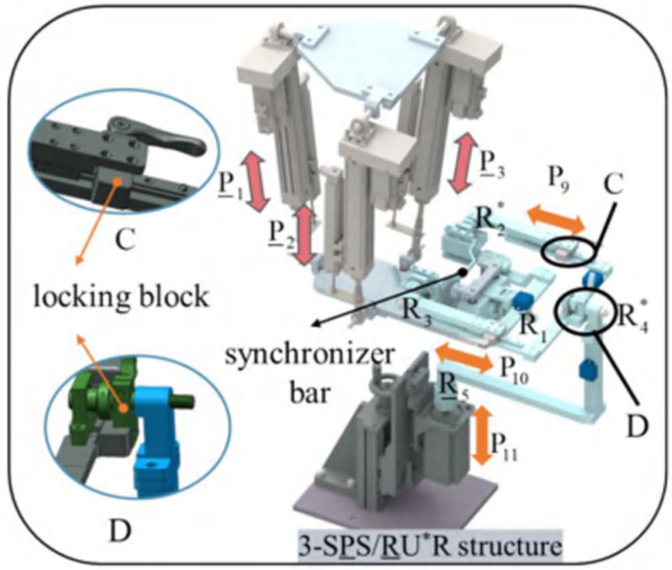


Figure 7. 3-SPS/RU*R parallel ankle rehabilitation robot. This figure is cited from [22].

Through mechanical connections such as gears and synchronous belts, the output is linearly related to the input signal, and the electromagnetic response time is extremely short. Consequently, motor control offers high precision and rapid response. However, motors are relatively heavy, have low power density, and often require a reducer. Common types include direct current (DC) motors, servo motors, and stepper motors. DC motors are suited for medium-precision, medium-load control; servo motors are preferred for high-precision, high-power applications; and stepper motors are ideal for low-speed, high-precision tasks.

Wu et al. developed an end-effector ankle rehabilitation robot using magnetic field–oriented control to drive a brushless DC motor, combined with planetary reduction gears of varying ratios to power all ankle motion actuators on a slider (**Figure 8**) [23].

Zeng et al. designed a rope-driven flexible ankle rehabilitation robot inspired by bionic principles [24]. The rope system, directly connected to the mobile platform via a steering pulley, is actuated by a servo motor. By employing a flexible equivalent motion axis and rope-drive mechanism that mimic the physiological ankle structure, this design addresses the human–machine motion mismatch seen in rigid ankle rehabilitation robots, enabling more natural movement compatibility (**Figure 9**).

Do et al. proposed a 1 DOF ankle rehabilitation system in which a stepper motor directly drives the pedals to perform dorsiflexion/plantarflexion [25]. This low-cost system aims to provide an economical and adaptable ankle rehabilitation solution for resource-limited settings (**Figure 10**).

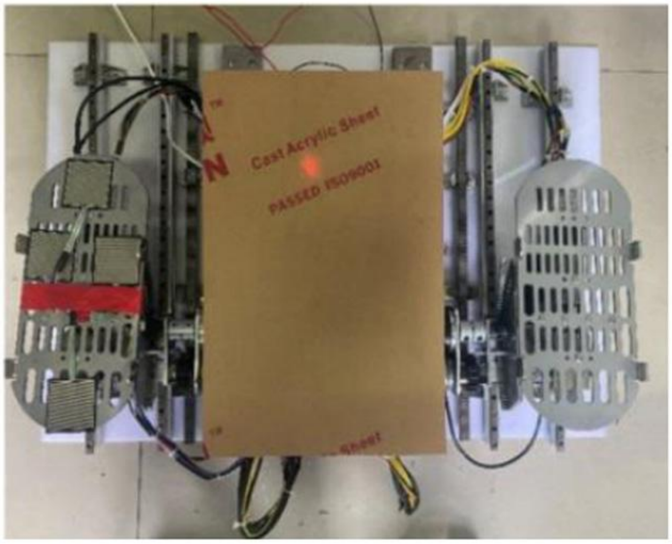


Figure 8. Brushless direct current motor ankle rehabilitation robot. This figure is cited from [23].

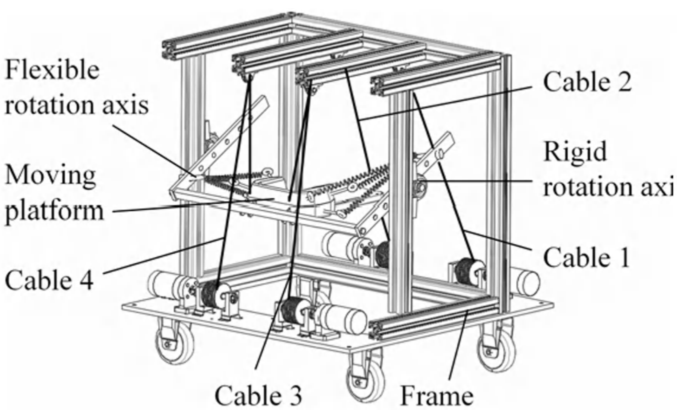


Figure 9. Servo motor ankle rehabilitation robot. This figure is cited from [24].

2.2.2 PMA

PMAs offer muscle-like compliance, making them well-suited for rehabilitation robots requiring flexibility and human-like motion, particularly in early rehabilitation stages or for compliant control. PMAs operate by using compressed air to contract a rubber or fabric tube, generating tension through material deformation. However, elastic hysteresis, creep, and the pressure transmission delay caused by air compression/expansion reduce control accuracy.

Lu et al. developed a flexible multi-DOF ankle rehabilitation robot using five PMAs, with tension transmitted to the moving platform via a flexible cable–pulley system [26]. A redundant drive configuration ensures forced closure and high platform stability (**Figure 11**).

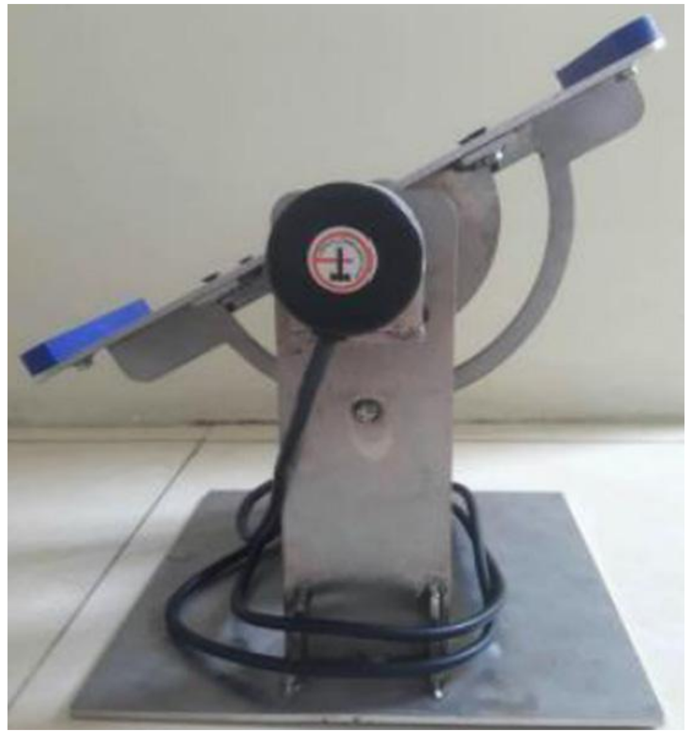


Figure 10. Stepper motor ankle rehabilitation robot. This figure is cited from [25].

Shamkhani et al. designed a wearable flexible ankle rehabilitation robot powered by six PMAs positioned around the foot (front, back, left, and right) [27]. Inflation and deflation of each PMA are manually controlled to provide assistance in all ankle joint directions (**Figure 12**).

Banyarani et al. proposed a soft wearable robot incorporating pleated fabric PMAs for walking assistance [28]. Their work covered mechanical modeling, manufacturing, exoskeleton system integration, biomechanical analysis, and experimental validation, demonstrating that pleated fabric PMAs offer lightweight, flexible support for individuals with disabilities or muscle weakness, reducing muscle fatigue during walking (**Figure 13**).

2.3 Ways to use

2.3.1 Platform ankle rehabilitation robot

Early ankle rehabilitation robots typically featured only 1-2 DOF and limited functionality. Since Girone et al. at Rutgers State University introduced the "Rutgers Ankle"—a 6 DOF PARR driven by pneumatic actuators based on a Stewart platform—the field has advanced significantly (**Figure 14**) [29]. The Stewart platform, widely used in flight simulators, precision positioning systems, robotics, and medical rehabilitation, comprises two parallel rigid platforms: a fixed lower platform and a movable upper platform with 6 DOF. The platforms are

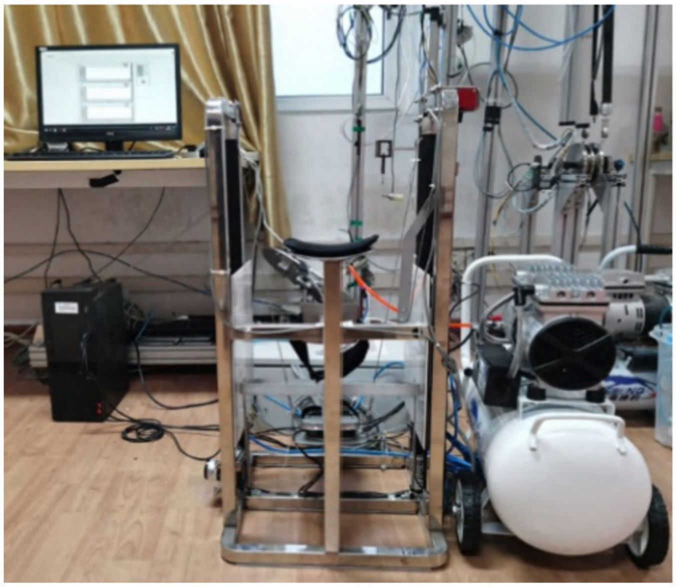


Figure 11. A flexible multi- degrees of freedom ankle rehabilitation robot. This figure is cited from [26].

Figure 12. A wearable flexible ankle rehabilitation robot. This figure is cited from [27].

connected by six independently adjustable retractable links, actuated pneumatically, hydraulically, or electrically, enabling translational and rotational motion of the upper platform in space.

Budaklı et al. developed a rehabilitation robot capable of simultaneous ankle and knee therapy by incorporating a seventh linear actuator and a specially designed mechanical structure, thereby expanding the traditional Stewart platform's workspace to accommodate knee rehabilitation (**Figure 15**) [30].

Wang et al. proposed a 6 DOF PARR employing flexible ball joints [31]. This design uses motor actuation, replaces conventional spherical joints with flexible hinge technology to eliminate clearance errors, and integrates Stewart platform architecture with compliant mechanism technology to form a novel hybrid structure (**Figure 16**).

Beyond Stewart-platform-based 6 DOF PARRs, other platform-type designs with varying actuator configurations and linkage arrangements are also applied. For instance, Zou et al. designed a 3-RRS PARR capable of performing both single-ankle and compound-ankle rehabilitation training [32].

2.3.2 Wearable ankle rehabilitation robot

Xia et al. developed a wearable rope-driven ankle exoskeleton capable of providing DO/PL assistance [33]. The system employs an iterative adaptive threshold algorithm combined with a peak detection algorithm to identify gait patterns in real time, enhancing detection accuracy. A hierarchical adaptive force–position hybrid control strategy is implemented to dynamically adjust control modes and optimize assistance. Preliminary tests on three healthy subjects demonstrated that the device effectively reduced metabolic cost during walking (**Figure 17**).

Wang et al. proposed a wearable ankle exoskeleton with a self-locking protection system based on plantar pressure feedback [34]. Real-time plantar pressure sensing is used to monitor gait state, distinguish normal walking from accidental falls, and improve fall detection accuracy through support vector machine–based data processing. The design integrates passive self-locking with active control strategies to provide dynamic ankle protection (**Figure 18**).

3 CONTROL STRATEGY DESIGN

Ankle rehabilitation robots employ various control methods, each of which can influence the patient's rehabilitation experience. Control strategies can be classified according to factors such as control objectives, interaction modes, and feedback signal types. For clarity, these strategies can be divided into two core categories: motion control and human–robot interaction control, corresponding to low-level and high-level control, respectively.

Low-level control focuses on the basic motion functions of ankle rehabilitation robots, including position control, speed control, and trajectory tracking. High-level control emphasizes interactive functions, with the most fundamental form being physical-layer interaction—such as force or resistance control. Building on this foundation, control can extend to intention-level interaction, where movement intention is recognized via bioelectrical signals such as electromyography (EMG) or electroencephalography (EEG), and to cognitive interaction, which

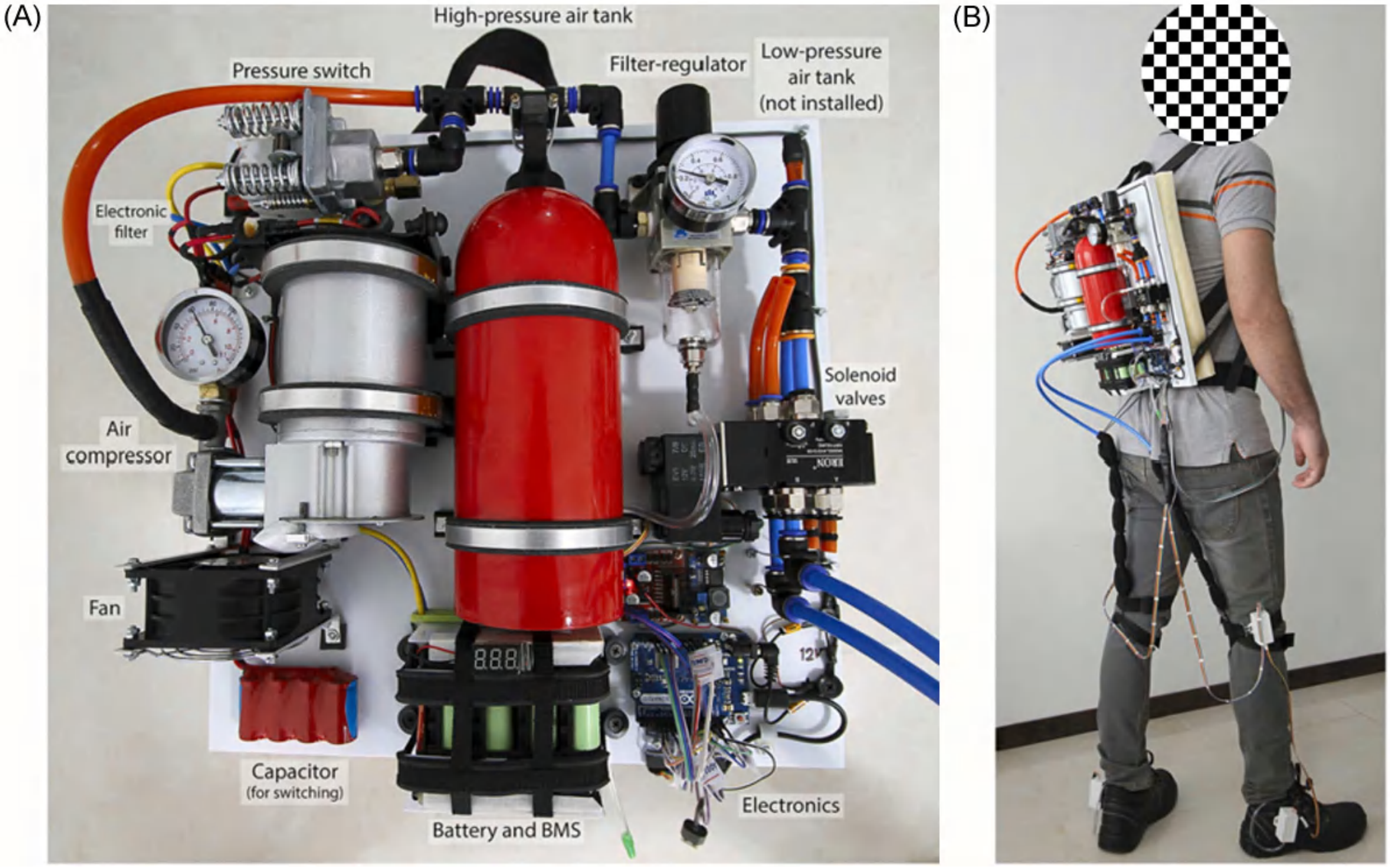


Figure 13. Wearable soft wearable robot. (A) Air supply backpack; (B) Wearing Diagram. This figure is cited from [28]. BMS, Battery Management System.

Figure 14. Rutgers Ankle. This figure is cited from [29].

uses multimodal control through VR or augmented reality (AR).

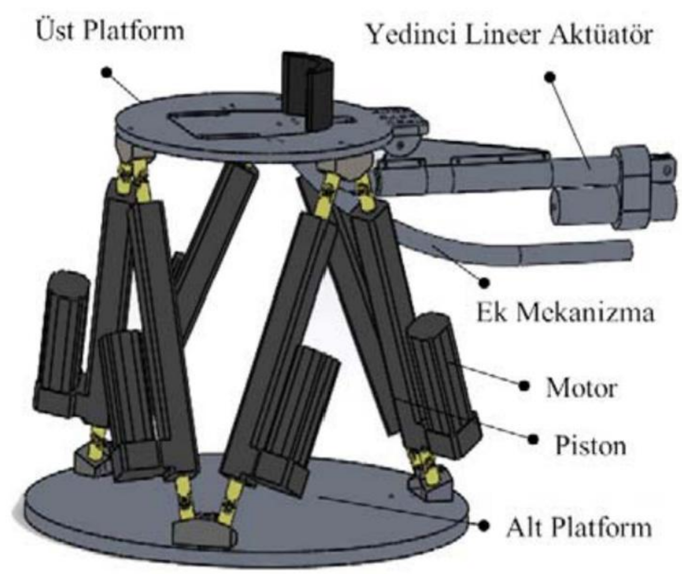


Figure 15. Ankle and knee rehabilitation robot based on Stewart. This figure is cited from [30].

Table 2 summarizes the design objectives, advantages, and limitations of different control methods—including position/speed control, trajectory tracking, force/impedance control,

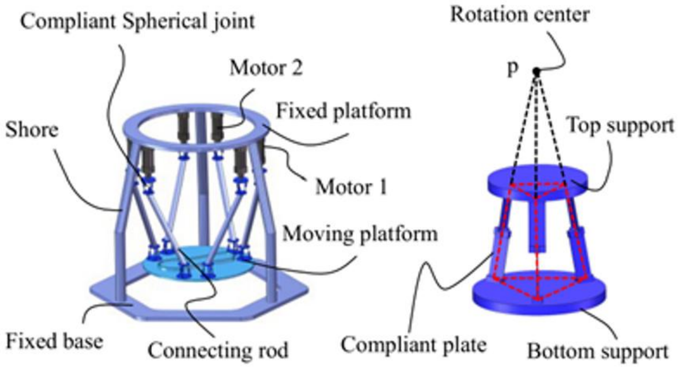


Figure 16. 6 DOF PARR based on flexible ball joints. This figure is cited from [31].

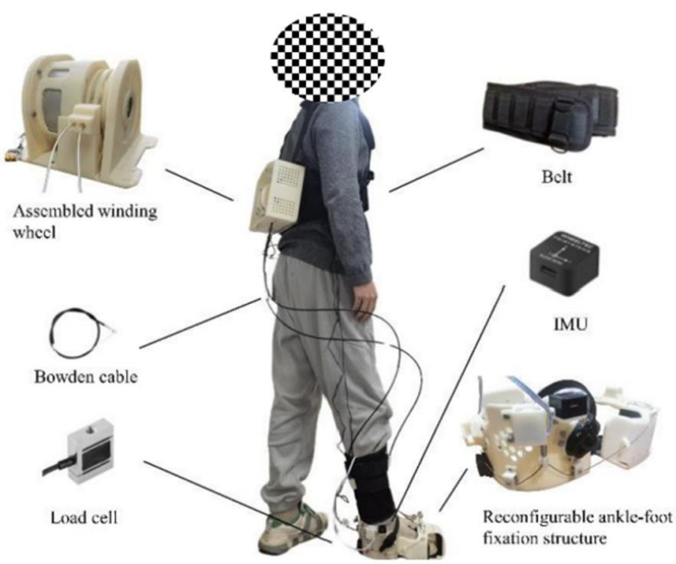


Figure 17. Wearable rope-driven ankle exoskeleton system. This figure is cited from [33]. IMU, Inertial Measurement Unit.

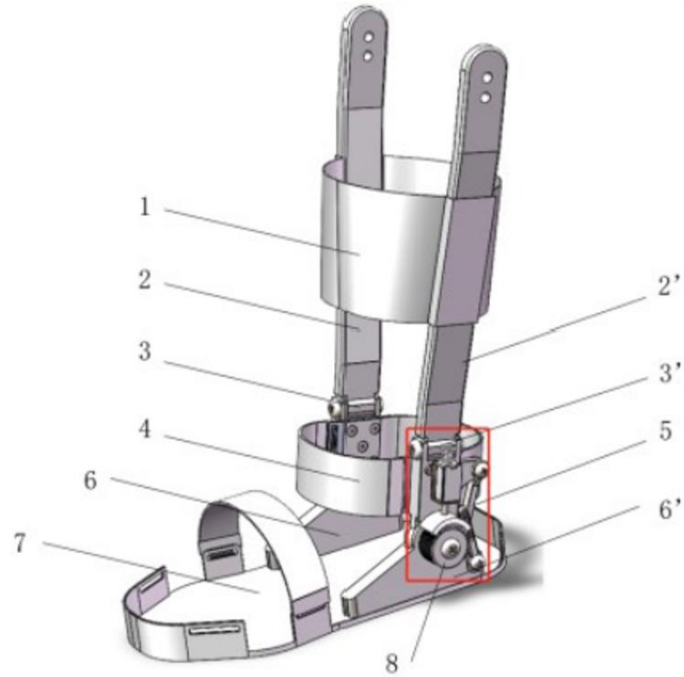


Figure 18. Wearable ankle exoskeleton self-locking protection system based on plantar pressure feedback. This figure is cited from [34]. 1, Fixing strap; 2, Leg support; 3, Joint bearing; 4, Leg support; 5, Achilles tendon connecting rod; 6, Support frame; 7, Base plate; 8, Self-locking device.

intention recognition, and cognitive interaction—according to their control strategy.

3.1 Motor function control

3.1.1 Position and speed control

Azizi et al. developed a DC motor position control system using a classical proportional–integral–derivative (PID) controller to regulate motor position and speed in ankle rehabilitation robots, enabling precise angle adjustment during training [35]. Lee et al. proposed a 2-DOF ankle exoskeleton employing PID neural network control, in which neural networks adaptively optimize PID parameters, providing high-precision and robust performance [36]. Li et al. introduced an isokinetic muscle training strategy based on adaptive gain and cascade PID control [37]. In this approach, when patients actively exert force, the robot maintains constant-speed motion while applying variable resistance, enhancing muscle training across different joint angles.

3.1.2 Trajectory control

Zhang et al. presented an adaptive trajectory tracking strategy for PARR systems with joint-space force distribution, combining cascade control with a motion-intention–guided algorithm to ensure accurate path following [38]. Liu et al. designed a multi-mode ankle rehabilitation robot for IN/EV training, normal gait training, and stepping training [39]. Dedicated trajectory plans for each mode enable complete motion execution and emergency stops at any position. Wang et al. developed a multi-stage hemiplegic lower-limb exoskeleton [40]. Gait data were collected via motion capture, calibrated with a backpropagation neural network, and fitted using Fourier functions to generate a high-precision trajectory model, achieving accurate multi-joint (hip, knee, ankle) tracking.

3.2 Human-robot interaction control

3.2.1 Force/torque control

Qian et al. proposed an iterative impedance learning method with a dual-loop structure: the outer loop adapts impedance parameters, while the inner loop executes torque control [41]. This allows patient-specific impedance adaptation through

Table 2. Summary and comparison of control strategies

Control strategy	Subcategories	Design purpose	Advantages	Limitation
Motion Control	Position/speed control	Achieve precise angle or speed adjustment	The algorithm is simple and reliable	Fixed parameters are difficult to adapt to individual differences or changes in rehabilitation progress
	Trajectory tracking control	Plan and track complex rehabilitation trajectories	Preset Trajectory, high stability	High computational complexity, relying on predefined trajectories or training data
Human-robot interaction control	Force/Impedance Control	Achieving safe human-machine interaction	The algorithm is simple and reliable	Sensitive to sensor accuracy and feedback delay
	Intention recognition control	Identifying patients' active movement intentions through bioelectric signals	Improving patient engagement and promoting neuroplasticity	Individual differences are large, and the signal is easily affected by interference
	Cognitive Interactive Control	Combine virtual scenes to enhance the fun of training	Enhance immersion, improve training fun and active participation	High hardware cost and complex algorithm development

repeated tasks, improving interaction quality. Liu et al. developed a redundant pneumatic muscle–cable-driven ankle robot with a hierarchical force–position control scheme [42]. Using the Karush–Kuhn–Tucker theorem and iterative optimization, force distribution is solved as a constrained problem, ensuring all actuators remain in tension for safety and controllability while combining precise trajectory tracking with safe force regulation. Li et al. introduced a dynamic admittance control strategy using a variable-operator fuzzy neural network to adjust virtual damping in real time, reducing muscle fatigue and extending training sessions according to patient needs [43].

3.2.2 Intention recognition control

Shi et al. presented an EEG-driven collaborative control model, extracting EEG deep features via convolutional neural networks and classifying them with a support vector machine to achieve high-precision intention recognition and active control, enhancing patient engagement [44]. Zhou et al. developed an sEMG-based ankle motion prediction model to estimate torque generation during movement, improving adaptability and control accuracy across individual differences [45].

3.2.3 Cognitive interaction control

Covaciu et al. designed a VR-based intelligent ankle rehabilitation system with an apple-catching game [46]. Motion data from accelerometers, gyroscopes, and EMG sensors are processed using a k-nearest neighbors algorithm to adapt task difficulty and control strategies, reducing therapist workload while enhancing engagement. Zhang et al. proposed a VR-assisted decision model based on a convolutional gated recurrent neural network, optimized via the whale optimization algorithm [47]. Immersive environments such as space walking and mountain climbing allow patients to control virtual characters through ankle movements for tasks like path following and obstacle avoidance. Vaida et al. developed an AR-integrated lower-limb rehabilitation system combining the LegUp parallel robot with HoloLens 2 [48]. Serious-game modules ("football mode" and "color mode") target motor coordination, strength, and cognitive training, offering multi-plane hip, knee, and ankle rehabilitation with real-time feedback.

4 CONCLUSION AND OUTLOOK

This review summarizes recent progress in the mechanical design and control strategies of ankle rehabilitation robots, highlighting notable advances while identifying persistent limitations. Series structures, though lightweight and suited for simple motion training, have limited dynamic performance, whereas parallel structures provide high precision and multi-DOF capability but require complex control, making them more appropriate for advanced rehabilitation and currently the most widely adopted. In actuation, motor drives offer high precision and strong rigidity for heavy-load applications but are less flexible, heavier, and less portable, while pneumatic muscles enable compliant control and passive stretching in early rehabilitation but suffer from nonlinear air pressure characteristics, inflation/deflation delays, and limited high-frequency responsiveness. Platform-based systems deliver stable support and accurate multi-DOF motion for early-stage passive training but are bulky and unsuitable for home use, whereas wearable devices, despite their portability, have lower load capacity and DOF, rely on batteries or pneumatic systems, and are better suited for mid- to late-stage gait recovery. Motion control strategies based on traditional robotics provide reliable and easy-to-implement algorithms for passive rehabilitation but are limited to basic movement execution, often failing to adapt to patient-specific needs or detect active intentions, which may reduce engagement. Human–robot interaction control, incorporating biological signals or VR/AR, enables more active participation, with intention recognition emerging as a key research focus; however, signal noise, individual variability, and real-time constraints remain technical challenges, and VR/AR integration, while improving engagement, increases system complexity and cost. Looking forward, the development of ankle rehabilitation robots is expected to move toward greater flexibility, intelligence, and accessibility by combining lightweight designs with flexible materials to enhance comfort, integrating multimodal biosignals such as EMG, EEG, plantar pressure, and inertial measurements into efficient AI-based control for improved intention recognition, and enriching VR/AR interaction through eye tracking, tactile feedback, and biomechanical sensing to create more immersive and realistic training. At the clinical level, adopting modular designs and scalable manufacturing will help reduce costs, enabling broader access to high-quality rehabilitation for patients in need.

DECLARATIONS

Author contributions

Jiajia Zha: Conceptualization, Methodology (Discussion), and Writing original draft preparation. Qingyun Meng: Conceptualization, Supervision, Project administration, and Writing reviewing and revising. Hongtao Shen and Mingxia Wei: Critical review and editing of the manuscript. All authors have read and agreed to the published version of the manuscript and ensure the integrity of the work.

Funding

This research received no external funding.

Data availability

All data analyzed during this study are included in this published article. No new data were created or analyzed.

Ethics approval and consent to participate

Not applicable. This is a literature review article that does not involve any new studies with human participants or animals.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

Not applicable.

REFERENCES

[1] Dong M, Zhou Y, Li J, Rong X, Fan W, Zhou X, et al. State of the art in parallel ankle rehabilitation robot: A systematic review. J Neuroeng Rehabil. 2021 Mar 20;18(1):52. https://doi.org/10.1186/s12984-021-00845-z

[2] Brockett CL, Chapman GJ. Biomechanics of the ankle. Orthop Trauma. 2016 Jun 15;30(3):232-238. https://doi.org/10.1016/j.mporth.2016.04.015

[3] Wells B, Allen C, Deyle G, Croy T. Management of acute grade II lateral ankle sprains with an emphasis on ligament protection: A descriptive case series. Int J Sports Phys Ther. 2019 Jun; 14(3):445-458. https://doi.org/10.26603/ijspt20190445

[4] Bonnel F, Toullec E, Mabit C, Tourné Y. Chronic ankle instability: Biomechanics and pathomechanics of ligaments injury and associated lesions. Orthop Traumatol Surg Res. 2010 Jun 1; 96(4):424-432. https://doi.org/10.1016/j.otsr.2010.04.003

[5] Yin Y, Lin Q, Wang J. Randomized controlled trial on ankle biomechanics in the treatment of functional ankle instability with joint mobilization. Sci Rep. 2024 Sep 27;14(1):22095. https://doi.org/10.1038/s41598-024-73646-8

[6] Hong R, Li B, Bao Y, Liu L, Jin L. Therapeutic robots for post-stroke rehabilitation. Med Rev. 2024 Feb 27;4(1):55-67. https://doi.org/doi:10.1515/mr-2023-0054

[7] Meng W, Zhu C, Liu H, Liu Q, Xie S. Soft ankle-foot exoskeleton for rehabilitation: A systematic review of actuation, sensing, mechanical design, and control strategy. IEEE Trans Med Robot Bionics. 2024 Apr 8;6(2):384-398. https://doi.org/10.1109/TMRB.2024.3385798

[8] Gribble PA, Kleis RE, Simon JE, Vela LI, Thomas AC. Differences in health-related quality of life among patients after ankle injury. Front Sports Act Living. 2022 Aug 3;4:909921. https://doi.org/10.3389/fspor.2022.909921

[9] Xing Y, Yang SD, Dong F, Wang MM, Feng YS, Zhang F. The beneficial role of early exercise training following stroke and possible mechanisms. Life Sci. 2018 Apr 1;198:32-37. https://doi.org/10.1016/j.lfs.2018.02.018

[10] Lee KE, Choi M, Jeoung B. Effectiveness of rehabilitation exercise in improving physical function of stroke patients: A systematic review. Int J Environ Res Public Health. 2022 Oct 5; 19(19):12739. https://doi.org/10.3390/ijerph191912739

[11] Dao MD, Tran XT, Pham DP, Ngo QA, Le TTT. Study on the ankle rehabilitation device. Arch Mech Eng. 2022 Dec 31; 69(1):147-163. https://doi.org/10.24425/ame.2021.139803

[12] Su F, Xu W. Enhancing brain plasticity to promote stroke recovery. Front Neurol. 2020 Oct 30;11:554089. https://doi.org/10.3389/fneur.2020.554089

[13] Devittori G, Dinacci D, Romiti D, Califfi A, Petrillo C, Rossi P, et al. Unsupervised robot-assisted rehabilitation after stroke: Feasibility, effect on therapy dose, and user experience. J Neuroeng Rehabil. 2024 Apr 9;21(1):52. https://doi.org/10.1186/s12984-024-01347-4

[14] Hau CT, Gouwanda D, Gopalai AA, Low CY, Hanapiah FA. Gamification and control of nitinol based ankle rehabilitation robot. Biomimetics. 2021 Sep 21;6(3):53. https://doi.org/10.3390/biomimetics6030053

[15] Nazar PS, Luebbe N, Pott PP. An affordable and safe system for game-based therapy of the ankle joint. Curr Dir Biomed Eng. 2023 Sep 20;9(1):270-273. https://doi.org/doi:10.1515/cdbme-2023-1068

[16] Ismail MAS, Adib MAHM. Rehabilitation and gamification technology device for lower extremities patient: A review. In: Hassan MHA, Omar MN, Johari NH, Zhong Y, Hassan MHA, Omar MN, Johari NH, Zhong YS, editors. Proceedings of the

2nd Human Engineering Symposium; 2023 Feb 14; Singapore. Springer Nature Singapore; 2023. p. 351-361.

[17] Lee S, Choi S, Ko C, Kim T, Kong K. Design and control of the compact cable-driven series elastic actuator module in soft wearable robot for ankle assistance. Int J Control Autom Syst. 2023 May 2;21(5):1624-1633. https://doi.org/10.1007/s12555-022-0098-4

[18] Eraky MT, Li A, Rocha MH, Teker A, Gebre BA, Nolan KJ, et al. A novel personalized ankle exoskeleton with co-located sea for gait training. 2024 10th IEEE RAS/EMBS International Conference for Biomedical Robotics and Biomechatronics 2024 Sep 1-4; Germany. IEEE; 2024. p. 1715-1720.

[19] Choi S, Kim J, Ko C, Kong K. Flexi-SEA: Flexible-shaft-driven series elastic actuator for wearable robots. IEEE ASME Trans Mechatron. 2025 Feb 26;30(6):6882-6892. https://doi.org/10.1109/TMECH.2025.3537113

[20] Wang F, Zheng Y, Hou Y, Wang Y, Xiao M. Design and analysis of parallel ankle joint rehabilitation robot. Int J Mechatron Appl Mech. 2024 Jun 14;(16):42-49. https://doi.org/10.17683/ijomam%2Fissue16.5

[21] Huo Y, Khan MN, Shao Z, Pan Y. Development of a novel cable-driven parallel robot for full-cycle ankle rehabilitation. Mechatronics. 2024 Aug 1;101:103210. https://doi.org/10.1016/j.mechatronics.2024.103210

[22] Xie Y, Zhong X, Wei J, Shi R. Kinematic performance analysis of a parallel ankle rehabilitation robot with sufficient distance adjustment. IJST-T Mech Eng. 2025 Apr 10;49(2):1109-1132. https://doi.org/10.1007/s40997-024-00829-z

[23] Wu R, Fan J, Wei J, Ma J, Zhang N, Luo M. Design of an end-effector ankle-foot rehabilitation robot with 3 DOF. 2023 7th International Conference on Robotics and Automation Sciences 2023 Jun 16-18; China. IEEE; 2023. p. 18-22.

[24] Zeng D, Liu Y, Qu C, Cong J, Hou Y, Lu W. Design and human-robot coupling performance analysis of flexible ankle rehabilitation robot. IEEE Robot Autom Lett. 2024 Nov 3;9(1):579-586. https://doi.org/10.1109/LRA.2023.3330052

[25] Do TV, Phuong NTL, Trang NTT, Lam LTH, Thanh PS, Duc NT. Ankle joint rehabilitation system—the preliminary results. Proceedings of the Sixth International Conference on Research in Intelligent and Computing; 2022 Jan 3-4; Vietnam. Polish Information Processing Society; 2022. p. 33-37.

[26] Lu Y, Zhang X, Ai Q. Mechanism design and analysis of a multi-DOF flexible ankle rehabilitation robot. J Phys Conf Ser. 2021 Apr;1884(1):012038. https://doi.org/10.1088/1742-6596/1884/1/012038

[27] Al-Shamkhani D, Al-Ibadi A, Giannaccini ME. Soft robot for ankle rehabilitation. 2023 16th International Conference on Developments in eSystems Engineering (DeSE); 2023 Mar 21; Turkiye. IEEE; 2023. p. 486-491.

[28] Banyarani PB, Tarvirdizadeh B, Hadi A. Design and fabrication of a soft wearable robot using a novel pleated fabric pneumatic artificial muscle (pfPAM) to assist walking. Sens Actuators A Phys. 2024 May;370:115278. https://doi.org/10.1016/j.sna.2024.115278

[29] Girone MJ, Burdea GC, Bouzit M. The “rutgers ankle” orthopedic rehabilitation interface. ASME 1999 International Mechanical Engineering Congress and Exposition; 1999 Nov 14-19; USA. American Society of Mechanical Engineers; 1999. p. 305-312.

[30] Budakli M, Yilmaz C. Stewart platform-based robot design and control for passive exercises in ankle and knee rehabilitation. J Fac Eng Archit Gazi Univ. 2021 Sep 2;36(4):1831-1846. https://doi.org/10.17341/gazimmfd.846641

[31] Wang H, Wang Q, Chen W, Yang Y, Li H. Design and research of a 6-degree-of-freedom parallel ankle rehabilitation robot based on compliant spherical joints. 2024 World Rehabilitation Robot Convention (WRRC); 2024 Aug 23-26; China. IEEE; 2024. p. 1-6.

[32] Zou Y, Zhang A, Zhang Q, Zhang B, Wu X, Qin T. Design and experimental research of 3-RRS parallel ankle rehabilitation robot. Micromachines. 2022 Jun 16;13(6):950. https://doi.org/10.3390/mi13060950

[33] Xia Y, Meng W, Ai Q, Zhu C, Liu Q. Design and control of wearable cable-driven ankle exoskeleton. 2024 43rd Chinese Control Conference (CCC); 2024 Jul 28-31; China. IEEE; 2024. p. 4749-4754.

[34] Wang D, Zheng Y, Wang F, Li C. Design and experimental validation of a wearable ankle exoskeleton with a self-locking mechanism based on plantar pressure. Int J Mech Appl Mech. 2025 Jan 6;1(19):122-132. https://doi.org/10.17683/ijomam/issue19

[35] Nutenki R, Varma BV. PID controller design for DC motor speed control. 2024 Fourth International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT); 2024 Jan 11-12; India. IEEE; 2024. p. 1-6.

[36] Lee T, Kim I, Baek YS. Design of a 2 DOF ankle exoskeleton with a polycentric structure and a bi-directional tendon-driven actuator controlled using a PID neural network. Actuators. 2021 Jan 4;10(1):9. https://doi.org/10.3390/act10010009

[37] Li J, Zhou Y, Dong M, Rong X. Isokinetic muscle strength training strategy of an ankle rehabilitation robot based on adaptive gain and cascade PID control. IEEE Trans Cogn Dev Syst. 2022 Jan 25;15(1):100-110. https://doi.org/10.1109/TCDS.2022.3145998

[38] Zhang M, McDaid A, Veale AJ, Peng Y, Xie SQ. Adaptive trajectory tracking control of a parallel ankle rehabilitation robot with joint-space force distribution. IEEE Access. 2019 Jun 26;7:85812-85820. https://doi.org/10.1109/ACCESS.2019.2925182

[39] Liu Y, Lu W, Wu H, Xia Y, Hu B, Zeng D. Performance analysis and trajectory planning of multi-locomotion mode ankle rehabilitation robot. Rob Auton Syst. 2022 Nov;157(2022):104246. https://doi.org/10.1016/j.robot.2022.104246

[40] Wang X, Wang H, Zhang B, Zheng D, Yu H, Cheng B, et al. A multistage hemiplegic lower-limb rehabilitation robot: Design and gait trajectory planning. Sensors. 2024 Apr 5;24(7):2310. https://doi.org/10.3390/s24072310

[41] Qian K, Zhang Z, Chakrabarty S, Xie S. Iterative impedance learning control for ankle rehabilitation. 2021 27th International Conference on Mechatronics and Machine Vision in Practice; 2021 Nov 26-28; China. IEEE; 2021. p. 492-497.

[42] Liu Q, Zuo J, Zhu C, Meng W, Ai Q, Xie SQ. Design and hierarchical force-position control of redundant pneumatic muscles-cable-driven ankle rehabilitation robot. IEEE Robot Autom Lett. 2022 Oct 29;7(1):502-509. https://doi.org/10.1109/LRA.2021.3123747

[43] Li J, Zhou Y, Zuo S, Dong M. Dynamically modulated robot compliance via online fuzzy neural networks for individualized ankle rehabilitation. Appl Intell. 2025 Feb 6;55(6):411. https://doi.org/10.1007/s10489-025-06317-4

[44] Shi M, Yang C, Zhang D. A novel human-machine collaboration model of an ankle joint rehabilitation robot driven by EEG signals. Math Probl Eng. 2021 Mar 12;2021(1):5564235. https://doi.org/10.1155/2021/5564235

[45] Zhou Y, Li J, Zuo S, Zhang J, Dong M, Sun Z. An online estimating framework for ankle actively exerted torque under multi-DOF coupled dynamic motions via sEMG. IEEE Trans Neural Syst Rehabil Eng. 2025 Dec 11;33:81-91. https://doi.org/10.1109/TNSRE.2024.3515966

[46] Covaciu F, Pisla A, Iordan AE. Development of a virtual reality simulator for an intelligent robotic system used in ankle rehabilitation. Sensors. 2021 Feb 23;21(4):1537. https://doi.org/10.3390/s21041537

[47] Zhang H, Liao Y, Zhu C, Meng W, Liu Q, Xie SQ. VR-aided ankle rehabilitation decision-making based on convolutional gated recurrent neural network. Sensors (Basel). 2024 Oct 30;24(21):6998. https://doi.org/10.3390/s24216998

[48] Vaida C, Rus G, Tucan P, Machado J, Pisla A, Zima I, et al. Enhancing robotic-assisted lower limb rehabilitation using augmented reality and serious gaming. Appl Sci. 2024 Dec 23;14(24):12029. https://doi.org/10.3390/app142412029

Progress in
Medical Devices

Zentime

RESEARCH ARTICLE

Design and analysis of a tissue retraction manipulator for neuroendoscopic surgery

Yu Liu, Gengqiang Shi

School of Health Sciences and Engineering, University of Shanghai for Science and Technology, Shanghai 200082, China.

Corresponding author: Gengqiang Shi.

Address correspondence to: Gengqiang Shi, School of Health Sciences and Engineering, University of Shanghai for Science and Technology, No. 334, Jungong Road, Yangpu District, Shanghai 200082, China.
E-mail: gengersgq@163.com.

Received October 24, 2025; Accepted January 8, 2026; Published March 24, 2026

DOI: 10.61189/091501wgyqdc

Abstract

This paper presents the design and analysis of a compact, cable-driven manipulator specifically for tissue retraction during neuroendoscopic surgery. The manipulator features an underactuated mechanism with a three-joint serial configuration, enabling stable motion within a single plane. Its compact design facilitates seamless integration into standard neuroendoscopic working channels, thereby optimizing spatial efficiency. The kinematic model was established using the Denavit-Hartenberg parameter method, with both forward and inverse kinematics systematically derived. Furthermore, a statics model was developed based on the Lagrangian formulation. Workspace analysis and trajectory planning were performed using Monte Carlo simulations in MATLAB. The simulation results indicate that the manipulator exhibits a feasible crescent-shaped workspace ($X \in [10, 50.9]$ mm, $Y \in [5.3, 44.9]$ mm). The motion trajectories of all joints were observed to be continuous and smooth, without any abrupt changes. Subsequent validation through ADAMS simulations confirmed the smooth variation of joint torques. This study provides a theoretical foundation and offers practical insights for the development and precise control of specialized instruments for neuroendoscopic surgery.

Keywords: Neuroendoscopic manipulator, Cable drive, Kinematic analysis, Simulation

1 INTRODUCTION

Neuroendoscopic surgery, a representative technique in modern minimally invasive neurosurgery, plays a vital and increasingly important role in treating intracranial lesions, owing to its significant advantages such as minimal trauma, rapid recovery, and short hospital stays [1, 2]. This technique involves advancing an endoscope and its working channel through narrow natural body cavities or micro-cranial bone windows to reach deep intracranial target areas, thereby providing surgeons with internal illumination and real-time visualization while minimizing damage to normal tissue [3, 4]. Precise and safe retraction or displacement of soft tissues—such as nerves and blood vessels—that obstruct the surgical field is critical for exposing key operative areas, preventing accidental injury, and ensuring surgical success [5]. However, neuroendoscopic surgery necessitates multi-handed manipulation by the surgeon. The highly minimally invasive nature of the procedure imposes severe spatial constraints, leading to significant interference between instruments [6]. Specifically, the manipulation of multiple instruments through the same channel frequently results in collisions, which severely compromises operational efficiency and safety [7, 8]. Surgeons therefore rely on handheld mechanical instruments to achieve stable and controllable tissue retraction or elevation.

To overcome the limitations of instruments in confined spaces, surgical robotics is considered a highly promising solution [9-11]. Commercialized laparoscopic surgical robots, such as the da Vinci SP system, demonstrate superior performance in larger cavities like the abdominal cavity through the introduction of wristed instruments [12-15]. However, their relatively large size makes them less suitable for minimally invasive procedures under strict spatial constraints. Consequently, research

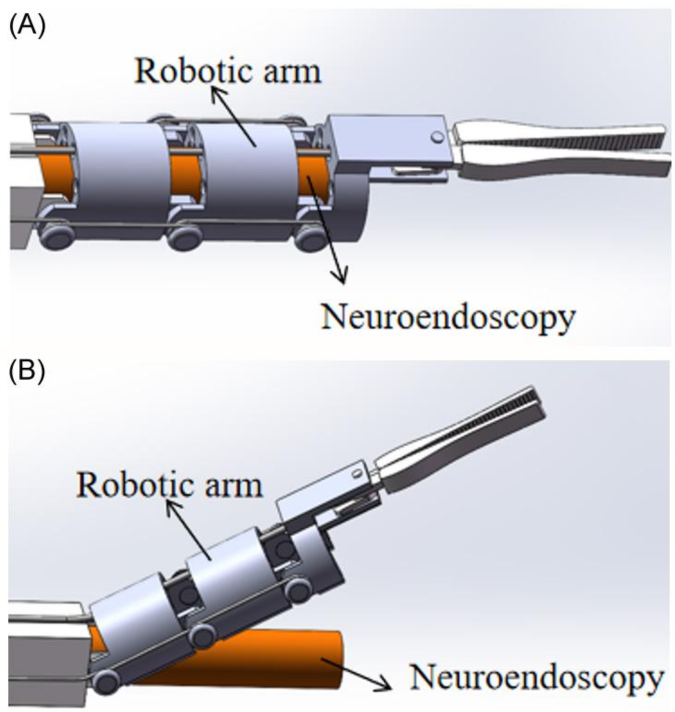


Figure 1. Schematic diagram of the mechanical hand structure model. (A) Retracted state; (B) Equivalent configuration during traction.

in this field has focused on highly flexible continuum robot designs [16-18].

Research efforts have explored various designs. For instance, a team from Incheon National University in South Korea developed a memory alloy-driven micro-flexible surgical robot [19]. Researchers at Shandong University developed a cable-driven surgical puncture robot, while Chinese scholars Luo et al. investigated a pneumatic/wire-driven continuum surgical robot for natural orifice surgery [20, 21]. A team from the Harbin Institute of Technology designed a magnetically driven millimeter-scale continuum robot for human lumen surgery [22]. A key challenge in the current research landscape is balancing the dexterity, size, stiffness, drive efficiency, and control complexity of such instruments. Practical devices specifically tailored for human body channels have yet to achieve significant breakthroughs. Among these diverse technical approaches, cable-driven systems are predominant in ultra-compact surgical instrument design due to their highly miniaturizable structure, immunity to electromagnetic interference, and ability to transmit power over long distances [23, 24]. However, many existing systems prioritize a high number of degrees of freedom, which often leads to excessive complexity without optimizing for core tasks such as planar tissue retraction in neuroendoscopic surgery. Building upon prior research, this paper introduces a specialized, compact, cable-driven manipulator specifically designed for planar tissue retraction in neuroendoscopic surgery. The Denavit–Hartenberg (D-H) parameter method and Monte Carlo simulations were employed to analyze the manipulator's forward and inverse kinematics and its effective workspace, thereby validating the rationality and effectiveness of the overall design. Motion trajectory planning simulations conducted in MATLAB further verified the motion continuity of the manipulator.

2 OVERALL STRUCTURE OF THE MANIPULATOR

2.1 Cable tension relationship analysis

To ensure stability and motion continuity, the manipulator employs an underactuated design [25]. This design features a minimalist joint topology consisting of three rotational joints arranged in a serial linkage configuration: a proximal finger pitch joint, a middle finger pitch joint, and an end-effector pitch/gripping joint (which provides tissue grasping/release functionality). The joint motion is constrained to a single plane, maximizing spatial efficiency. This approach addresses the critical requirement for surgical field exposure in neuroendoscopy while avoiding the complexity and spatial inefficiency associated with redundant degrees of freedom. All joint motions are actuated by a compact cable-driven system, ensuring a simple mechanical structure that can be integrated directly into the standard working channel of a neuroendoscope. Pitch motion is generated by winding cables around fixed pulleys located on both sides of each joint. Grasping and releasing functions of the gripping joint are operated by cables pulled through the central aperture of the joint. The cables are guided from the proximal end and routed counterclockwise over the fixed pulleys to drive the three joints. The gripping joint is secured by winding cables around its dissection aperture, as shown in **Figure 1**.

The rotation angle of each joint is determined by the length of cable being pulled. In the neutral position, the length of cable wrapped around each fixed pulley equals the pulley's circumference. During rotation, the angular displacement corresponds to the arc length of the cable released, as shown in **Figure 2**. The relationship between joint angle and cable length is established through the following geometric derivation:

$$S = R \cdot \theta_1 \tag{1}$$

Applying the arc angle conversion formula yields:

$$\theta_2 = \theta_1 \cdot \frac{180}{\pi} \tag{2}$$

Where θ_1 is the angle of arc displacement, S is the arc length, R is the radius of the fixed pulley, and θ_2 is the angle of the center circle.

Consequently, the relationship between the center circle angle and cable length variation is expressed as:

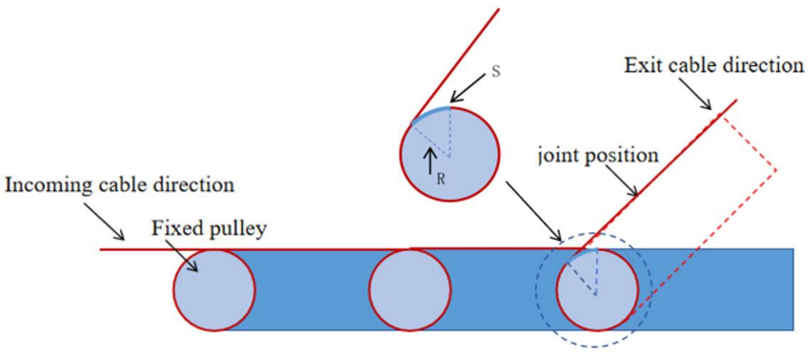


Figure 2. Schematic diagram of the cable joint tension relationship.

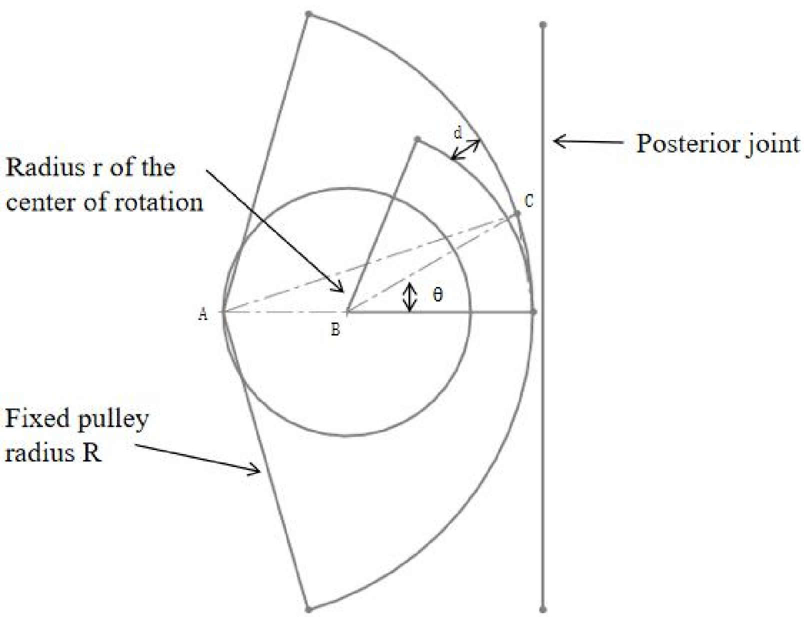


Figure 3. Schematic diagram of the design of the maximum angle of the joint.

$$S = R \cdot \theta_2 \cdot \frac{\pi}{180} \tag{3}$$

2.2 Maximum joint angle analysis

The maximum rotation angle θ for each joint is designed to be 30°, representing the ideal range of motion obtained through kinematic calculation. Due to an offset between the center of the arc at the joint end and the center of the fixed pulley, a specific clearance distance maintained between adjacent joints leads to a displacement during rotation. When this displacement equals the clearance at the rear joint, the two joints undergo rigid collision, halting rotation.

As shown in **Figure 3**, by constructing triangle ABC and applying trigonometric relations, the following is obtained:

$$AC^2 = AB^2 + BC^2 - AB \cdot BC \cdot \cos(\pi - \theta) \tag{4}$$

Solving equation (4) yields:

$$R^2 = (R - r)^2 + (r + d)^2 - (R - r) \cdot (r + d) \cdot \cos(\pi - \theta) \tag{5}$$

Equation (5) indicates that for given radii of the joint arc and the fixed pulley, the maximum angle θ before collision occurs is governed by the clearance distance d. Thus, the maximum angle θ can be optimally designed by appropriately adjusting the clearance distance.

3 KINEMATIC ANALYSIS OF THE MANIPULATOR

The kinematics of the three finger joints, which bend under cable tension, are analyzed using the standard D-H parameter method to model the relationship between their range of motion and configuration changes. The parameter coordinate systems between the manipulator's joints are shown in **Figure 4**. The D-H parameters and corresponding joint ranges of motion for the manipulator are listed in **Table 1**.

3.1 Forward kinematics

Forward kinematics provides the foundation for robotic motion control and path planning, involving the mathematical derivation of the end-effector pose from a given set of joint angles. It establishes the transformation between the robot's base frame and the end-effector frame using a 4×4 homogeneous transformation matrix. This section presents forward and inverse kinematic analyses of the manipulator based on the established D-H parameter model [26]. The overall transformation matrix is obtained by successively multiplying the transformation matrices of individual joints.

The transformation matrix between adjacent robot links is defined as:

$${}^{\,i}_{i+1}T = Rot(z, \theta_{i+1}) \times Trans(0, 0, d_{i+1}) \times Trans(a_{i+1}, 0, 0) \times Rot(x, \alpha_{i+1}) \tag{6}$$

$$= \begin{bmatrix} C\theta_{i+1} & -S\theta_{i+1}C\alpha_{i+1} & S\theta_{i+1}S\alpha_{i+1} & \alpha_{i+1}C\theta_{i+1} \\ S\theta_{i+1} & C\theta_{i+1}C\alpha_{i+1} & -C\theta_{i+1}S\alpha_{i+1} & \alpha_{i+1}S\theta_{i+1} \\ 0 & S\alpha_{i+1} & C\alpha_{i+1} & d_{i+1} \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{7}$$

Where the following notation is used: cosθ=Cθ, sinθ=Sθ; cosα=Cα, sinα=Sα.

The joint rotation matrices, starting from the proximal joint, are as follows:

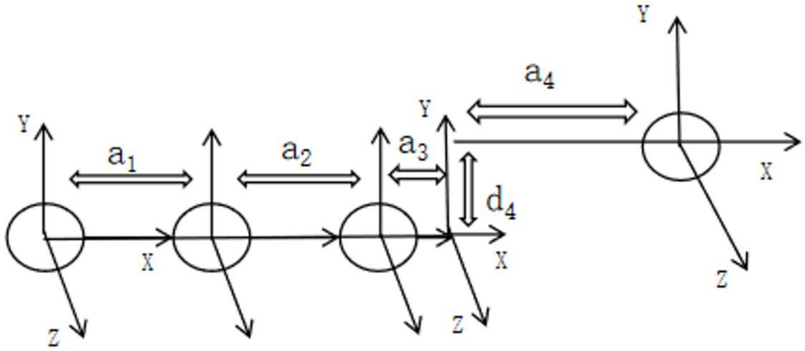


Figure 4. Denavit–Hartenberg coordinate diagram of the manipulator.

Table 1. Denavit–Hartenberg parameters of the manipulator

i	θ	d(mm)	α	a(mm)
1	$\theta_1(0,30^0)$	0	0	$a_1(10)$
2	$\theta_2(0,30^0)$	0	0	$a_2(10)$
3	$\theta_3(0,30^0)$	0	0	$a_3(5)$
4	$\theta_4(0)$	$d_4(4.5)$	0	$a_4(25)$

$$
{}^0_1T = \begin{bmatrix} C\theta_1 & -S\theta_1 & 0 & a_1 C\theta_1 \\ S\theta_1 & C\theta_1 & 0 & a_1 S\theta_1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{8}
$$

$$
{}^1_2T = \begin{bmatrix} C\theta_2 & -S\theta_2 & 0 & a_2 C\theta_2 \\ S\theta_2 & C\theta_2 & 0 & a_2 S\theta_2 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{9}
$$

$$
{}^2_3T = \begin{bmatrix} C\theta_3 & -S\theta_3 & 0 & a_3 C\theta_3 \\ S\theta_3 & C\theta_3 & 0 & a_3 S\theta_3 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{10}
$$

The final homogeneous transformation matrix defines the pose of the end-effector's tip (gripping point) relative to the base frame:

$$
{}^3_4T = \begin{bmatrix} 1 & 0 & 0 & a4 \\ 0 & 1 & 0 & d4 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{11}
$$

Denoting the homogeneous transformation matrix of the end-effector gripping point as:

$$
{}^0_4T = \begin{bmatrix} n_x & o_x & a_x & p_x \\ n_y & o_y & a_y & p_y \\ n_z & o_z & a_z & p_z \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{12}
$$

Based on the D-H method, overall transformation matrix for the end effector pose is obtained by concatenating the individual joint transformations:

$$
{}^0_4T = {}^0_1T \times {}^1_2T \times {}^2_3T \times {}^3_4T \tag{13}
$$

The resulting elements of ${}^0_4T$ are:

$$
\begin{aligned}
&n_x = C\theta_{123}, o_x = -S\theta_{123}, a_x = 0, \\
&p_x = a_1 C\theta_1 + a_2 C\theta_{12} + a_3 C\theta_{123} + a_4 C\theta_{123} - d_4 S\theta_{123}, \\
&n_y = S\theta_{123}, o_y = C\theta_{123}, a_y = 0, \\
&p_y = a_1 S\theta_1 + a_2 S\theta_{12} + a_3 S\theta_{123} + a_4 S\theta_{123} + d_4 C\theta_{123}, \\
&n_z = 0, o_z = 0, a_z = 1, p_z = 0,
\end{aligned}
$$

Where the abbreviated notation $C\theta_{ij}=cos(\theta_i+\theta_j)$, $S\theta_{ij}=sin(\theta_i+\theta_j)$ is used.

3.2 Inverse kinematics

The inverse kinematics analysis determines the rotational angles θ of each joint based on the homogeneous transformation matrix of the robot's end-effector joint pose, whose elements are given in Equation (12). This section uses the inverse transformation method to solve for the joint angles. The solution procedure is outlined as follows:

From Equation (13), we obtain:

$$
\theta_{123} = \arctan\frac{n_y}{n_x} \tag{14}
$$

Order:

$$
p_x - a_3 n_x - a_4 n_x + d_4 n_y = a_1 C\theta_1 + a_2 C\theta_{12} = K_x \tag{15}
$$

$$
p_y - a_3 n_x - a_4 n_x - d_4 n_y = a_1 S\theta_1 + a_2 S\theta_{12} = K_y \tag{16}
$$

Available:

$$
\begin{aligned}
K_x^2 + K_y^2 &= a_1^2 + a_2^2 + 2a_1 a_2 C\theta_2 \\
&= (p_x - a_3 n_x - a_4 n_x + d_4 n_y)^2 + (p_y - a_3 n_x - a_4 n_x - d_4 n_y)^2
\end{aligned} \tag{17}
$$

Then the equation for θ_2 is:

$$
\theta_2 = arccos\frac{K_x^2 + K_y^2 - a_1^2 - a_2^2}{2a_1 a_2} \tag{18}
$$

Rewriting Equation (15)-(16) as:

$$K_x - a_1 C\theta_1 = a_2 C\theta_{12}$$
$$K_y - a_1 S\theta_1 = a_2 S\theta_{12} \tag{19}$$

Squaring both sides of the equation and adding yields:

$$K_x C\theta_1 + K_y S\theta_1 = \frac{K_x^2 + K_y^2 + a_1^2 - a_2^2}{2a_1} \tag{20}$$

Define

$$\cos\varnothing = \frac{K_x}{\sqrt{K_x^2 + K_y^2}} \tag{21}$$

$$\sin\varnothing = \frac{K_y}{\sqrt{K_x^2 + K_y^2}} \tag{22}$$

So that

$$\varnothing = arctan\frac{K_y}{K_x} \tag{23}$$

Using trigonometric relationships, the expression for θ_1 is:

$$\sqrt{K_x^2 + K_y^2}\,(cos\varnothing cos\theta_1 + sin\varnothing sin\theta_1) = \frac{K_x^2 + K_y^2 + a_1^2 - a_2^2}{2a_1} \tag{24}$$

$$\theta_1 = \varnothing + arccos\frac{\frac{K_x^2 + K_y^2 + a_1^2 - a_2^2}{2a_1}}{\sqrt{K_x^2 + K_y^2}} \tag{25}$$

Then the equation for θ_3 is:

$$\theta_3 = arctan\frac{n_y}{n_x} - \theta_1 - \theta_2 \tag{26}$$

3.3 Static analysis

Based on the Lagrangian formulation for serial robotic arms, the dynamic equations of the finger mechanism are expressed as:

$$D(q)\ddot{q} + C(q,\dot{q}) + G(q) + \tau_f(\dot{q}) = \tau_t \tag{27}$$

Where $q=[\theta_1\ \theta_2\ \theta_3]^T$, $q,\dot{q},\ddot{q}$ denote the angular displacement, velocity, and acceleration vectors of the finger joints; $D(q)$ is the inertia matrix; $C(q,\dot{q})$ represents the centrifugal and Coriolis torque terms; $G(q)$ is the gravitational torque; $\tau_f(\dot{q})$ denotes joint friction torque; τ_t is the joint driving torque.

The elements of these matrices are calculated based on the geometric and inertial parameters of the three links. Substituting these elements into Equation (27) yields the complete dynamic equations of the finger mechanism.

The three finger segments are denoted as link 1, link 2, and link 3, with masses m_1, m_2, m_3, and moments of inertia about their respective centers of mass.

The Lagrange method is employed to solve for the joint torques. The fundamental expression of the Lagrange equation is:

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_i}\right) - \frac{\partial L}{\partial q_i} = \tau_t, i = 1,2,3 \tag{28}$$

In the equation, L represents the Lagrangian operator, where L=K-P, with K the total kinetic energy and P the total potential energy of the linkage system. To proceed, the linear velocities, accelerations, and angular accelerations of each link must be determined individually. Substituting the Lagrangian into equation (27) and performing differential calculations yields the corresponding torque matrix.

For link 1, the position of its center of mass in the coordinate system $O_1\ X_1Y_1Z_1$ can be represented as ${}^1p_{r1}=[r_1,0,0]^T$.

Rotating about the origin of the basis coordinate system with angular velocity $\dot{\theta}_1$, its linear velocity is $v_1 = r_1\dot{\theta}_1$.

The kinetic and potential energies of link 1 are respectively:

$$K_1 = \frac{1}{2}m_1 v_1^2 + \frac{1}{2}I_1\dot{\theta}_1^2 = \frac{1}{2}(m_1 r_1^2 + I_1)\dot{\theta}_1^2 \tag{29}$$

$$P_1 = m_1 g_1 r_1 S\theta_1 \tag{30}$$

For rod 2, its center of mass in the coordinate system $O_2\ X_2Y_2Z_2$ can be expressed as ${}^2p_{r_2} = [r_2, 0, 0]^T$.

Through coordinate transformation, its position ${}^0p_{r_2}$ in the base coordinate system can be expressed as:

$${}^0p_{r_2} = {}^0A_1{}^1A_2{}^2p_{r_2} = \begin{bmatrix} l_1 C\theta_1 + r_2 C\theta_{12} \\ l_1 S\theta_1 + r_2 S\theta_{12} \\ 0 \end{bmatrix} \tag{31}$$

According to equation (31), the absolute velocity of the center of mass of rod 2 can be calculated as:

$$\frac{d{}^0p_{r_2}}{dt} = \begin{bmatrix} -l_1\dot{\theta}_1 S\theta_1 - r_2\dot{\theta}_{12} S\theta_{12} \\ l_1\dot{\theta}_1 C\theta_1 + r_2\dot{\theta}_{12} C\theta_{12} \\ 0 \end{bmatrix} \tag{32}$$

Accordingly, the kinetic energy K_2 of rod 2 can be determined as:

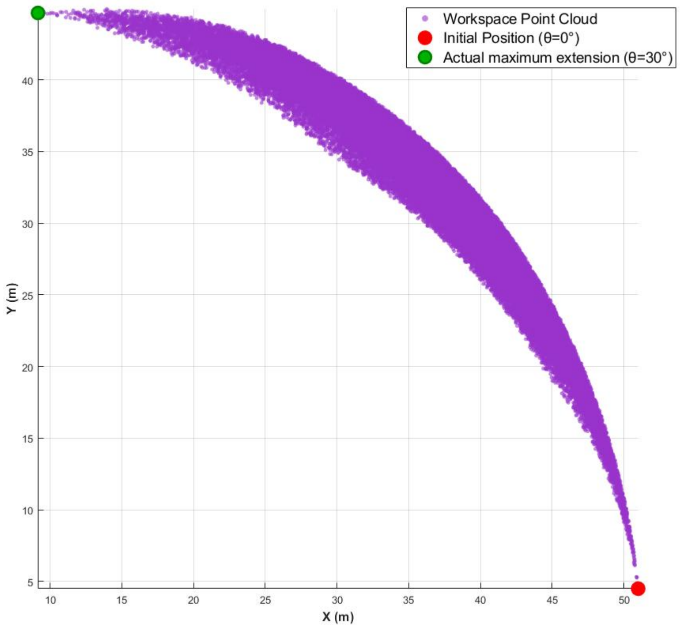


Figure 5. Workspace schematic of the robotic hand.

$$K_2 = \frac{1}{2} m_2 v_2^2 + \frac{1}{2} I_2 \dot{\theta}_2^2 \tag{33}$$

The linear velocity v_2 is:

$$v_2^2 = l_1^2 \dot{\theta}_1^2 + r_2^2 \dot{\theta}_{12}^2 + 2 l_1 r_2 \dot{\theta}_1 \dot{\theta}_{12} C\theta_2 \tag{34}$$

Potential energy P_2 is:

$$P_2 = m_2 g_2 \left[l_1 S\theta_1 + r_2 S\theta_{12} \right] \tag{35}$$

For rod 3, its center of mass in the coordinate system O_3-$X_3Y_3Z_3$ can be expressed as ${}^3p_{r_3}=[r_3,h_3,0]^T$.

Through coordinate transformation, its position ${}^0p_{r_2}$ in the base coordinate system can be expressed as:

$$ {}^0pr_3 = {}^0A_1 {}^1A_2 {}^2A_3 {}^3pr_3 = \begin{bmatrix} l_1 C\theta_1 + l_2 C\theta_{12} + r_3 C\theta_{123} - h_3 C\theta_{123} \\ l_1 S\theta_1 + l_2 S\theta_{12} + r_3 S\theta_{123} + h_3 C\theta_{123} \\ 0 \end{bmatrix} \tag{36}$$

According to equation (36), the absolute velocity of the center of mass of rod 2 can be calculated as:

$$\frac{d\,{}^0pr_3}{dt} = \begin{bmatrix} -l_1 \dot{\theta}_1 S\theta_1 - l_2 \dot{\theta}_{12} S\theta_{12} - r_3 \dot{\theta}_{123} S\theta_{123} - h_3 \dot{\theta}_{123} C\theta_{123} \\ l_1 \dot{\theta}_1 C\theta_1 + l_2 \dot{\theta}_{12} C\theta_{12} + r_3 \dot{\theta}_{123} C\theta_{123} - h_3 \dot{\theta}_{123} S\theta_{123} \\ 0 \end{bmatrix} \tag{37}$$

Accordingly, the kinetic energy K_3 of rod 3 can be determined as:

$$K_3 = \frac{1}{2} m_3 v_3^2 + \frac{1}{2} I_3 \dot{\theta}_3^2 \tag{38}$$

The linear velocity v_3 is:

$$\begin{aligned} v_3^2 &= l_1^2 \dot{\theta}_1^2 + l_2^2 \dot{\theta}_{12}^2 + (r_3^2 + h_3^2) \dot{\theta}_{123}^2 + 2 l_1 l_2 \dot{\theta}_1 \dot{\theta}_{12} C\theta_2 \\ &+ 2 l_1 r_3 \dot{\theta}_1 \dot{\theta}_{123} C\theta_{23} - 2 l_1 h_3 \dot{\theta}_1 \dot{\theta}_{123} S\theta_{23} \\ &+ 2 l_2 r_3 \dot{\theta}_{12} \dot{\theta}_{123} C\theta_3 - 2 l_2 h_3 \dot{\theta}_{12} \dot{\theta}_{123} S\theta_3 \end{aligned} \tag{39}$$

Potential energy P_3 is:

$$P_3 = m_3 g_3 \left[l_1 S\theta_1 + l_2 S\theta_{12} + r_3 S\theta_{123} + h_3 C\theta_{123} \right] \tag{40}$$

Both the kinetic energy and potential energy of the linkage system have been determined, enabling the calculation of the Lagrangian operator:

$$L = K - P = K_1 + K_2 + K_3 - P_1 - P_3 - P_3 \tag{41}$$

Using the Lagrange operator L, calculate the partial derivatives for each joint variable: $\frac{\partial L}{\partial \theta_i}$, $\frac{\partial L}{\partial \dot{\theta}_i}$, $\frac{d}{dt}\frac{\partial L}{\partial \dot{\theta} i}$, i=1,2,3.

Substituting the above results into the Lagrange equations (27) yields the joint torques.

4 ROBOTIC ARM SIMULATION ANALYSIS

The workspace of the manipulator was analyzed using the Monte Carlo method [27]. A large number of random joint angle combinations were generated within their allowable ranges, and the corresponding end-effector positions were computed via the forward kinematics model in MATLAB. The point cloud representing these positions was plotted to visualize the workspace. As shown in **Figure 5**, the robotic arm's motion range forms a crescent shape, with spatial extents of X∈(10, 50.9) mm and Y∈(5.3, 44.9) mm. This extensive motion range meets the practical movement requirements of each mechanical joint and effectively fulfills the design task of expanded traction.

Trajectory planning is a critical aspect of robotics, as it significantly influences the performance and smoothness of robotic motion. In this study, a point-to-point trajectory between specified initial and terminal poses was planned. The trajectory was generated using quintic polynomial interpolation via the MATLAB Robotics Toolbox, based on the established kinematic model. This process generated the angle, angular velocity, and angular acceleration profiles for the three joints along the planned path.

Starting pose:

$$\begin{bmatrix} 1 & 0 & 0 & 51.00 \\ 0 & 1 & 0 & 4.50 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{42}$$

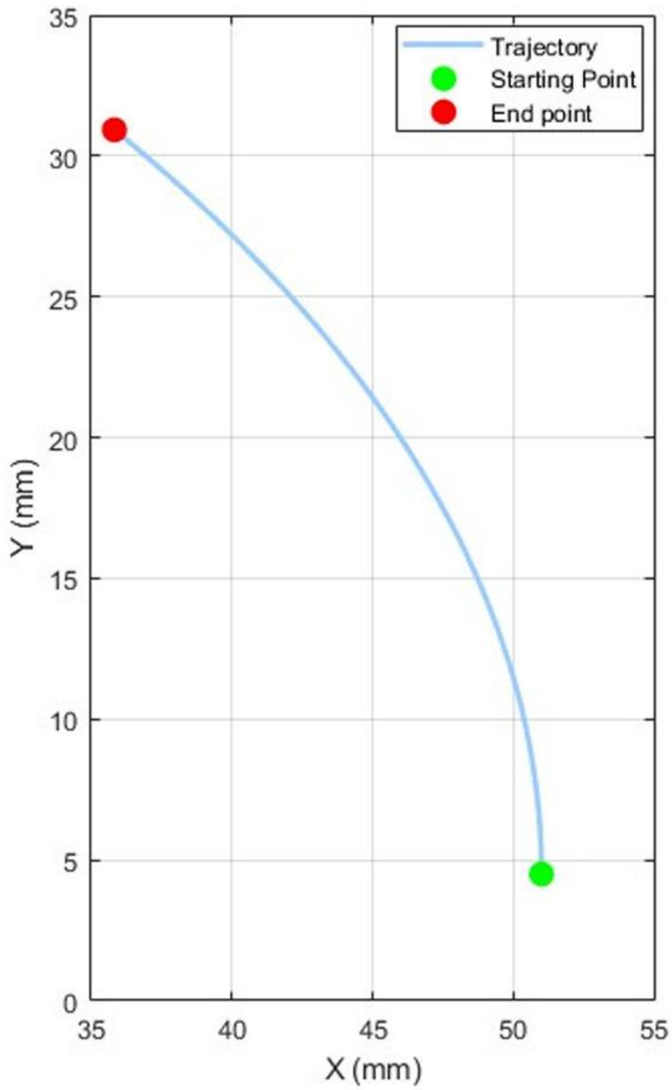


Figure 6. Trajectory planning results.

Terminal pose:

$$\begin{bmatrix} 0.642788 & -0.766044 & 0 & 35.84 \\ 0.766044 & 0.642788 & 0 & 30.93 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{43}$$

The resulting trajectories are shown in **Figure 6**.

As illustrated in **Figures 7-9**, the angular displacement, velocity, and acceleration profiles of all three joints demonstrate smooth transitions without abrupt changes during motion along the planned trajectory. This finding indicates that all joints can be driven stably throughout the workspace, thereby validating the rationality of the mechanical design. The manipulator model developed in SOLIDWORKS was subsequently imported into ADAMS for dynamic simulation. Within the ADAMS environment, the material properties were defined as steel, and the corresponding joint torque profiles were acquired. Collectively, these simulation results provide an intuitive and practical reference for the subsequent development of the physical platform and the design of its control system.

As shown in **Figure 10**, the negative values on the Y-axis indicate only the direction of the torque, not its magnitude. The torque variation curve reveals that the torque at all three joints exhibits an approximately S-shaped continuous trend. The transitions are smooth and uniform, with no abrupt changes observed. This indicates that the load transition at each joint during motion is stable, meeting the design expectations and stability requirements.

5 CONCLUSION

This study addresses the practical need for tissue retraction in neuroendoscopic surgery by presenting the design of a compact, cable-driven manipulator tailored for minimally invasive procedures. The proposed manipulator employs an underactuated design with a minimalist three-joint topology, achieving stable planar motion. This configuration significantly reduces the spatial footprint while ensuring full functionality, thereby enabling seamless integration into standard neuroendoscopic working channels. This makes the manipulator particularly suitable for confined surgical environments.

In terms of mechanism design, the maximum joint rotation angle was mechanically constrained through the coordinated design of joint arcs and fixed pulleys, coupled with precise clearance control. A kinematic model was established based on the D-H parameter method, and the forward and inverse kinematic relationships were systematically derived. Furthermore, a static model was constructed using the Lagrange method, providing a theoretical foundation for control and force feedback. Trajectory planning analysis conducted in MATLAB demonstrated that the manipulator possesses a feasible crescent-shaped workspace, which is sufficient for the required traction and exposure of surgical sites. The combined results from the trajectory and torque simulations revealed continuous and smooth motion profiles for all joints, with no abrupt changes, further validating the rationality of the mechanism design and its motion stability.

While this study has made significant progress in structural design and kinematic analysis, several limitations remain to be addressed. For instance, the current model does not incorporate effects such as time delay and deformation caused by cable flexibility. Furthermore, the mechanical performance under extreme loading conditions requires further experimental vali-

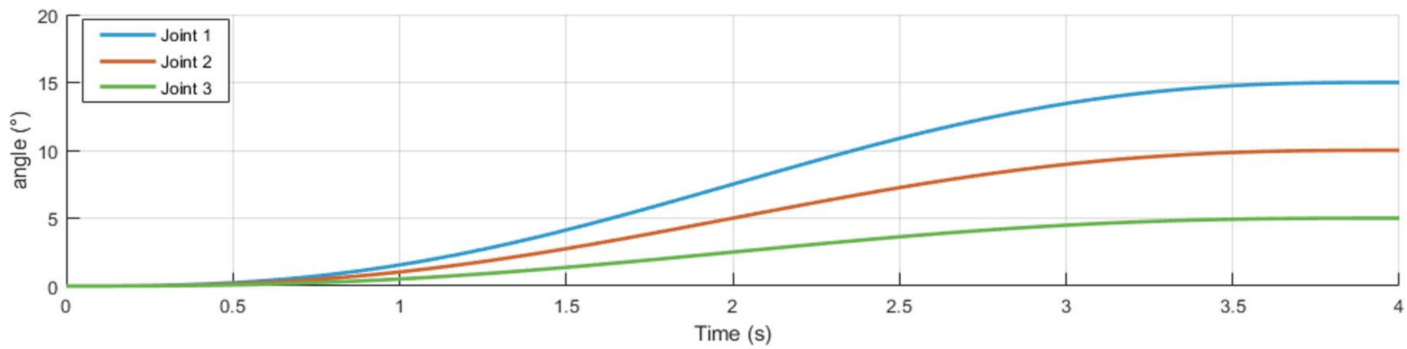


Figure 7. Joint angle profiles.

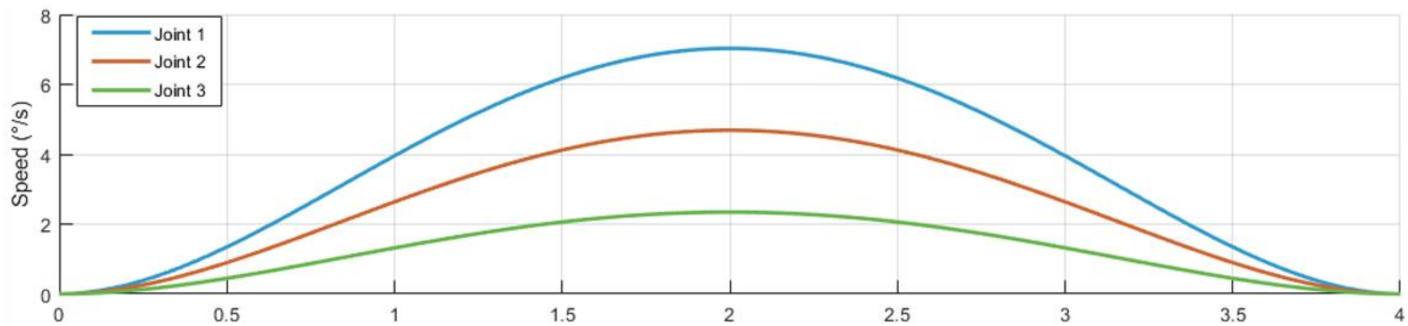


Figure 8. Joint angular velocity profiles.

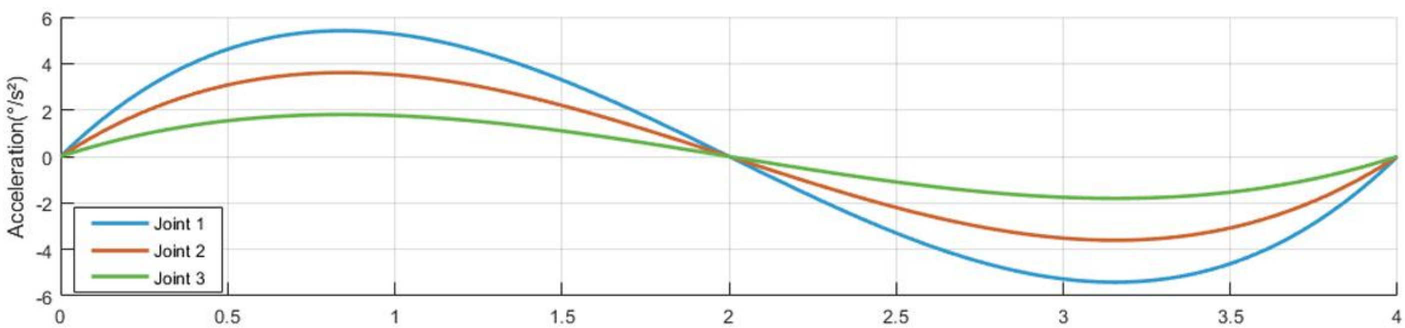


Figure 9. Joint angular acceleration profiles.

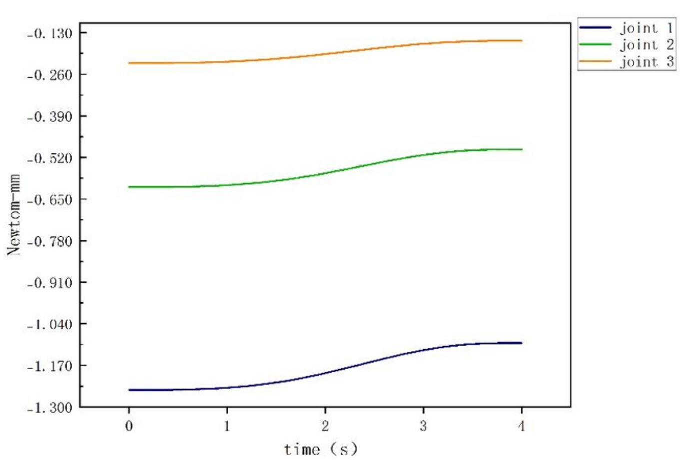


Figure 10. Joint torque profiles.

dation. Future work also includes coordinating the manipulator's motion with a robotic arm to achieve free movement in three-dimensional space, which warrants further investigation.

DECLARATIONS

Author contributions

Both authors contributed equally to the work. Liu participated in the implementation and drafting of the manuscript, while Shi contributed to the conceptual development and design of the study.

Funding

This research received no external funding.

Data availability

The data presented in this study are authentic and reliable.

Ethics approval and consent to participate

Not applicable (this study did not involve any procedures requiring ethical review).

Consent for publication

The submitted manuscript has not been published elsewhere (except in the form of an abstract or a conference presentation) and is not currently under consideration for publication by any other journal.

Competing interests

The authors declare no potential conflicts of interest with respect to the research, authorship, or publication of this article.

Acknowledgements

Not applicable.

REFERENCES

[1] Xu L, Lu X, Zhang C, Wang W. Clinical efficacy of neuroendoscopy combined with intracranial pressure monitoring for the treatment of hypertensive intracerebral hemorrhage. World Neurosurg. 2024 Jul;187:e210-e219. https://doi.org/10.1016/j.wneu.2024.04.068

[2] Banoci J, Magocova V, Katuch V, Tomkova M. Neuroendoscopy-assisted evacuation for supratentorial intracerebral hemorrhage versus conventional craniotomy: A comparative analysis of efficacy and outcome. Bratisl Lek Listy. 2024;125(12):775-779. https://doi.org/10.4149/bll_2024_118

[3] Wang L, Zhou T, Wang P, Zhang S, Yin Y, Chen L, et al. Efficacy and safety of neuroendoscopic surgery for intracerebral hemorrhage: A randomized, controlled, open-label, blinded endpoint trial (nesich). Int J Stroke. 2024 Jun;19(5):587-592. https://doi.org/10.1177/17474930241232292

[4] Zhou L, Jing X, Wang C, Zhang H, Lei P, Song P, et al. Clinical application of transcranial neuroendoscopy combined with supraorbital keyhole approach in minimally invasive surgery of the anterior skull base. Sci Rep. 2024 Jun 27;14(1):14886. https://doi.org/10.1038/s41598-024-65758-y

[5] Spennato P, Tengattini F, Picariello S, Gritti L, Porzio M, Ruggiero C, et al. The role of neuroendoscopy in treatment of pediatric brain abscesses: Case series and systematic review of the literature. Neurosurg Rev. 2025 Jun 20;48(1):518. https://doi.org/10.1007/s10143-025-03669-w

[6] Zhou L, Xu G, Liu K, Zhang H, Lei P, Lu M, et al. 3d slicer combined with neuroendoscopic surgery for the treatment of basal ganglia hemorrhage after cranioplasty: A case report and literature review. Heliyon. 2024 Sep 30;10(18):e37773. https://doi.org/10.1016/j.heliyon.2024.e37773

[7] Monteiro GA, Marinheiro G, Mutarelli A, Araújo B, Cavalcante-Neto JF, Batista S, et al. Efficacy and safety of neuroendoscopy surgery versus craniotomy for supratentorial intracerebral hemorrhage: An updated meta-analysis of randomized controlled trials. Neurosurg Rev. 2024 Jun 4;47(1):255. https://doi.org/10.1007/s10143-024-02492-z

[8] Waseem MH, Abideen ZU, Rehman N, Khan MH, Tahir MF, Raja HAA, et al. Efficacy and safety of neuroendoscopy versus craniotomy for spontaneous supratentorial intracerebral hemorrhage: An updated meta-analysis of randomized and non-randomized studies. Brain Behav. 2025 Sep;15(9):e70581. https://doi.org/10.1002/brb3.70581

[9] Omisore OM, Han S, Xiong J, Li H, Li Z, Wang L. A review on flexible robotic systems for minimally invasive surgery. IEEE Trans Syst Man Cybern Syst. 2022;52(1):631-644. https://doi.org/10.1109/TSMC.2020.3026174

[10] Shkurti TE, Çavuşoğlu MC. A systematic review of task automation in surgical robotics. IEEE Trans Med Robot Bionics. 2025;7(3):863-880. https://doi.org/10.1109/TMRB.2025.3583182

[11] Li C, Zhang G, Zhao B, Xie D, Du H, Duan X, et al. Advances of surgical robotics: Image-guided classification and application. Natl Sci Rev. 2024 Sep;11(9):nwae186. https://doi.org/10.1093/nsr/nwae186

[12] Park JI, Chung YK, Lee YM, Nam CW, Nah YW. Comparative analysis of postoperative outcomes of single-incision cholecystectomy: Propensity score matching of robotic surgery using the da Vinci SP system and da Vinci Xi system vs. laparoscopic surgery. Ann Hepatobiliary Pancreat Surg. 2025 May 31;29(2):140-149. https://doi.org/10.14701/ahbps.24-198

[13] Miyamura H, Takada K, Ohwaki A, Kobayashi A, Ito M, Nomura H, et al. Initial experience and surgical outcomes of robotic-assisted total hysterectomy using the da Vinci SP surgical system. Asian J Endosc Surg. 2024 Apr;17(2):e13298. https://doi.org/10.1111/ases.13298

[14] Heo JE, Han HH, Lee J, Choi YD, Jang WS. Single-port robot-assisted pyeloplasty using the da Vinci SP system versus multi-port pyeloplasty: Comparison of outcomes and costs. Asian J Surg. 2024 Sep;47(9):3841-3846. https://doi.org/10.1016/j.asjsur.2024.03.175

[15] Vizza R, Corrado G, Mancini E, Baiocco E, Russo M, Vincenzoni C, et al. Feasibility, safety, and efficacy of robotic single-port hysterectomy (R-SPH) using the da Vinci SP system in low-risk endometrial cancer: A pilot study. Minim Invasive Ther Allied Technol. 2025 Dec;34(6):433-440. https://doi.org/10.1080/13645706.2025.2538764

[16] Wang Z, Weng D, Li Z, Chen L, Ma Y, Wang J. A magnetic-controlled flexible continuum robot with different deformation modes for vascular interventional navigation surgery. Actuators. 2023;12(6):247. https://doi.org/10.3390/act12060247

[17] Troncoso DA, Robles-Linares JA, Russo M, Elbanna MA, Wild S, Dong X, et al. A continuum robot for remote applications: From industrial to medical surgery with slender continuum robots. IEEE Robot Autom Mag. 2023;30(3):94-105. https://doi.org/10.1109/MRA.2022.3223220

[18] Sun L, Chen X. Flexible continuum robot system for minimally invasive endoluminal gastrointestinal endoscopy. Machines. 2024;12(6):370. https://doi.org/10.3390/machines12060370

[19] Jeon H, Le QN, Jeong S, Jang S, Jung H, Chang H, et al. Towards a snake-like flexible robot with variable stiffness using an sma spring-based friction change mechanism. IEEE Robot Autom Lett. 2022;7(3):6582-6589. https://doi.org/10.1109/LRA.2022.3174363

[20] Li J, Jiang Q. Optimal design and experiment of cable-driven puncturing surgery robot for soft needle. J Med Device. 2023 Feb 8;17:1-37. https://doi.org/10.1115/1.4056865

[21] Luo X, Song D, Zhang Z, Wang S, Shi C. A novel distal hybrid pneumatic/cable-driven continuum joint with variable stiffness capacity for flexible gastrointestinal endoscopy. Adv Intell Syst. 2023 Jun 1;5(6):2200403. https://doi.org/10.1002/aisy.202200403

[22] Mao L, Yang P, Tian C, Shen X, Wang F, Zhang H, et al. Magnetic steering continuum robot for transluminal procedures with programmable shape and functionalities. Nat Commun. 2024 May 4;15(1):3759. https://doi.org/10.1038/s41467-024-48058-x

[23] Zhang F, Fu Y, Yang L, Fu Y. A novel cable configuration method for fully-actuated parallel cable-driven systems: Application in a shoulder rehabilitation exoskeleton. Mech Mach Theory. 2024 Sep 1;199:105693. https://doi.org/10.1016/j.mechmachtheory.2024.105693

[24] Qin G, Cheng Y, Ji A, Pan H, Yang Y, Yao Z, et al. Research on the cable-driven endoscopic manipulator for fusion reactors. Nucl Eng Technol. 2024 Feb 1;56(2):498-505. https://doi.org/10.1016/j.net.2023.10.026

[25] Wu X, Li X, Li Z, Zhang D, Miao Z, Zhou J. An improved udwadia–kalaba approach for controller design in underactuated mechanical systems. Multibody Syst Dyn. 2025 Jun 1;64(2):195-218. https://doi.org/10.1007/s11044-024-10004-6

[26] Jia L, Liu S, Cao C, Kang Y, Zhu Y, Wang L, et al. Kinematics and spatial structure analysis of TBM gunite robot based on D-H parameter method. Sci Rep. 2024 Jun 13;14(1):13649. https://doi.org/10.1038/s41598-024-64439-0

[27] Yang Z, Tian W, Wang H, Liu X, Zhang D, Yan Y, et al. Snake-like robot workspace solving method based on improved monte carlo method. Appl Bionics Biomech. 2025;2025:6125695. https://doi.org/10.1155/abb/6125695

Progress in
Medical Devices

Zentime

RESEARCH ARTICLE

Heart sound classification based on the fusion of dynamic features and images of mel-frequency cepstral coefficients

Shoucheng Chen, Rongguo Yan, Ke Wang, Wenjing Du

School of Biomedical Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China.

Corresponding author: Rongguo Yan.

Address correspondence to: Rongguo Yan, School of Health Science and Engineering, University of Shanghai for Science and Technology, No. 334, Jungong Road, Shanghai 200093, China. E-mail: yanrongguo@usst.edu.cn.

Received October 25, 2025; Accepted December 4, 2025; Published March 24, 2026

DOI: 10.61189/371147mjbess

Abstract

Heart sound analysis plays a key role in the early screening and auxiliary diagnosis of cardiovascular diseases. However, conventional auscultation largely depends on physicians' personal experience, which often leads to subjective and inconsistent evaluations. To overcome these limitations, this paper presents an intelligent heart sound classification framework that integrates dynamic mel-frequency cepstral coefficient (MFCC) features with dynamic MFCC-based images. In this work, the static MFCCs together with their first- and second-order derivatives are extracted to describe both the spectral and temporal behaviors of heart sounds. A multi-branch fusion model is designed to enhance feature interaction among the dynamic MFCC features via cross-branch attention. Meanwhile, a CA-ResNet18 network incorporating a coordinate attention mechanism is employed to learn spatio-temporal representations from the dynamic MFCC images. The high-level features produced by both models are then concatenated and classified using a support vector machine. Experimental validation on the PhysioNet Challenge 2016 dataset demonstrates that the proposed method achieves 96.82% accuracy, 97.51% sensitivity, and 96.19% specificity. Comparative studies with recent state-of-the-art methods confirm that the proposed integration of dynamic feature fusion and hybrid deep learning–machine learning framework significantly enhances the robustness and classification performance in intelligent heart sound analysis.

Keywords: Heart sound classification, Dynamic MFCC, Multi-branch fusion, Coordinate attention, Support vector machine

1 INTRODUCTION

The 2024 China Cardiovascular Health and Disease Report pointed out that from 1990 to 2019, the incidence of cardiovascular disease (CVD) in China increased by 132.82% over that 30-year period [1]. CVDs are characterized by acute onset and high severity, which usually quickly lead to life-threatening events. Therefore, the early prevention and accurate diagnosis of CVDs are of great clinical significance.

Acoustic auscultation is a generally accepted initial screening method among the currently available diagnostic methods, which is non-invasive, convenient and cost-effective. Its clinical significance stems from the association between organic heart disease and abnormal heart sounds. The analysis of acoustic signals has reference value in the early detection and risk stratification of CVDs. However, traditional auscultation has limitations. Its effectiveness relies heavily on the experience of clinicians and involves subjective interpretation of findings, which lacks objectivity and standardisation. Consequently, these limitations hinder the screening of the whole population and compromise both the consistency and accuracy of diagnosis.

With the continuous and rapid development of digital signal processing and artificial intelligence technology, computer-aided diagnosis systems provide a new way to mediate the limitations of traditional auscultation. Through digital processing of sound signals, the computer-aided diagnosis system can automatically identify the typical characteristics of the sample and use intelligent learning algorithms to further improve the standardisation and reliability of its voice processing. Subsequently, this can improve the accuracy in detecting abnormalities and the consistency of clinical diagnosis.

The general process of heart sound classification includes three stages: preprocessing, feature extraction, and classification model training. Preprocessing is mainly used for denoising and segmentation, while feature extraction and model design are key to achieving high classification performance. The current methods for classifying heart sounds can be roughly divided into three categories: traditional methods based on manual features, methods based on deep learning, and methods based on multi-feature fusion.

Early research relied heavily on manually designed features combined with traditional classifiers.

For example, Li et al. used wavelet transform and Twin Support Vector Machine to achieve heart sound recognition; Xu et al. integrated time-domain and frequency-domain features with random forest and AdaBoost to classify congenital heart disease in children; Taneja et al. used gammatonegram features combined with a support vector machine (SVM) to achieve heart sound classification [2-4]. Although these methods can achieve certain results, they are difficult to capture complex temporal patterns.

With the development of deep learning, researchers have begun to use neural networks to automatically extract high-level features. Rubin et al. combined mel-frequency cepstral coefficients (MFCCs) with convolutional neural networks (CNNs); Deng et al. proposed a Convolutional Recurrent Network based on improved MFCC; Shahid et al. used spectrograms combined with CNN and SVM for heart sound classification; Xiao et al. and Oh et al. directly used the original heart sound signal as input and implemented end-to-end classification using an attention mechanism or WaveNet [5-9]. Li et al. developed the Multiscale DenseNet and Multi-head Attention Recurrent Neural Network (MDN-MARNN) model, which uses a multi-scale DenseNet structure and a multi-head self-focus RNN, which is an end-to-end classification method based on error label data with a certain degree of interpretability [10]. Patwa et al. proposed a composite 1D-CNN and wavelet scattering transformation (WST) model, which provides robustness under small sample testing and class imbalance [11]. Cheng et al. proposed the CNN-Transformer End-to-end Neural Network (CTENN) model, which utilizes CNN to extract local features and Transformer to capture global dependencies, achieving end-to-end heart sound classification and performing well on the PhysioNet dataset [12]. Compared to traditional methods, these models significantly improve feature learning ability and classification performance.

In recent years, multi-feature fusion has emerged as a new research focus. This type of method enhances the ability to characterize heart sound signals by integrating features of different types or levels. Abbas et al. fused spectrograms with MFCC to enhance robustness; Lee et al. combined wavelet transform and CNN to achieve multi-scale feature extraction; Li et al. proposed the CAFusionNet model, which significantly improves classification accuracy through multi-layer feature fusion and a channel attention mechanism [13-15]. Overall, the introduction of multi-feature fusion and attention mechanisms provides new research directions and performance breakthroughs for intelligent classification of heart sounds. Khan et al. proposed a heart sound spectrogram classification method based on residual learning and multi-model feature fusion [16]. MobileNet and DenseNet201 were used to extract features and residual fusion was used to enhance the model expression ability. The method achieved excellent heart rhythm anomaly detection performance on both PhysioNet 2016 and BreakHis datasets, and was combined with Gradient-weighted Class Activation Mapping to provide model interpretability.

In two-dimensional acoustic representation, MFCC features are still one of the most common features due to their perceptually motivated spectral representation. Nevertheless, traditional MFCCs only represent, i.e., characterise, the spectral envelope of a selected frame, and thus do not capture the temporal dynamics of heart sounds. In this study, dynamic MFCC features are derived to capture the temporal dynamics. Furthermore, these dynamic MFCCs are visualised to generate dynamic MFCC feature map. These dynamic MFCC feature maps are then used to construct and integrate two feature extraction models, the outputs of which are combined and classified using an SVM classifier. Consequently, the proposed approach improves the robustness and accuracy of heart sound classification by effectively combining temporal dynamic and spectral representations.

2 METHODS AND EQUIPMENT

2.1 Static MFCC

MFCC is one of the most common acoustic characteristics in biomedical signal and speech processing. It represents the perceptual characteristics of the human auditory system, and its frequency response is nonlinear, because people are more sensitive to low frequencies than to high frequencies. MFCC uses the Mel scale to model this perceptual nonlinearity, making it an appropriate representation of the spectral envelope of sound signals [17, 18]. The workflow of MFCC feature extraction is illustrated in **Figure 1**.

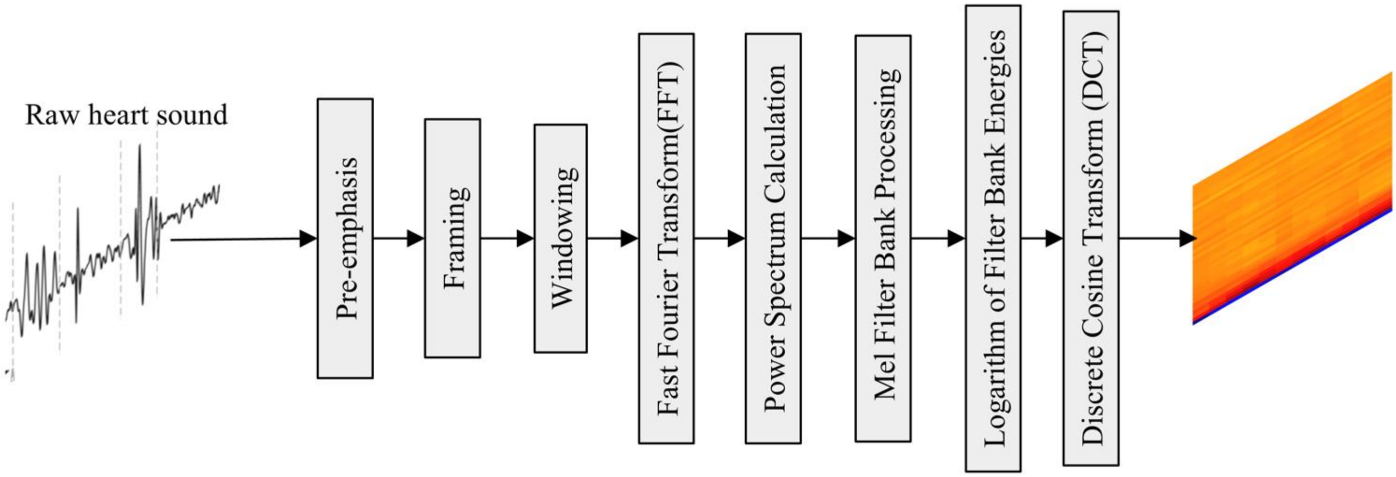


Figure 1. Workflow of MFCC feature extraction. MFCC, mel-frequency cepstral coefficient.

The pre-emphasis in the heart sound classification is to enhance the high-frequency component and reduce the low-frequency energy at the same time. This improves the detection of subtle changes and pathophysiological information that are usually embedded in high-frequency bands. Pre-emphasis is defined as the following equation:

$$y(n) = x(n) - ax(n-1) \tag{1}$$

Where $x(n)$ is the original signal, $y(n)$ is the pre-emphasis output, and the pre-emphasis coefficient α is usually between 0.9 and 1.0.

Heart sound is a quasi-stationary signal, which allows the original heart sound to be divided into short, overlapping time frames to enable spectral decomposition of each frame. In order to maintain continuity and reduce spectral leakage, each time frame is multiplied by a window function. In heart sound processing, the Hamming window and Hann window are usually used as window functions. Although their functions are similar, the Hamming window has a wider main lobe and better stopband roll-off, which allows for flexible suppression of side lobe influence. In contrast, the Hann window usually has better resolution and narrower side lobes (which also have lower amplitude) [19]. The Hamming window can be mathematically represented as:

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), 0 \le n \le N-1 \tag{2}$$

When segmenting the heart sound signal, parameters such as segment duration (t), sampling frequency (f_s), frame length (F_l), and frame shift (F_m) must be considered, as they determine the total number of frames M, given by:

$$M = \frac{t \times f_s - F_l}{F_m} + 1 \tag{3}$$

Since heart sound signals exhibit less distinct time-domain features, it is necessary to analyze their spectral content. A Fast Fourier Transform is applied to each windowed frame to convert the signal from the time domain to the frequency domain:

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-j\frac{2\pi kn}{N}, k=0,1,\ldots,N-1} \tag{4}$$

The power spectrum provides a quantitative measure of energy distribution across frequencies, reflecting cardiac activity intensity at each frequency component:

$$P(k) = \frac{1}{N}\left[X(k)\right]^2 \tag{5}$$

The Mel filter bank is designed to approximate the human auditory perception of frequency. The conversion from frequency f(Hz) to the Mel scale is expressed as:

$$mel(f) = 2595\log_{10}\left(1 + \frac{f}{700}\right) \tag{6}$$

The inverse conversion from Mel scale to frequency is given by:

$$f = 700\left(10^{\frac{mel}{2595}} - 1\right) \tag{7}$$

Each Mel filter has a triangular response, with a maximum at the center frequency and overlapping ranges between adjacent filters to ensure smooth spectral coverage. The transfer function of the m^{th} Mel filter is defined as:

$$H_m(k) = \begin{cases} 0 & , k < f(m-1) \\ \dfrac{k - f(m-1)}{f(m) - f(m-1)} & , f(m-1) \le k \le f(m) \\ \dfrac{f(m+1) - k}{f(m+1) - f(m)} & , f(m) < k \le f(m+1) \\ 0 & , k > f(m+1) \end{cases} \tag{8}$$

Where $f(m)$ denotes the center frequency of the m^{th} filter, $f(m-1)$ and $f(m+1)$ represent the lower and upper bounds, respectively. These frequencies are computed as:

$$f(m) = (\frac{S}{f_s}) F_{Mel}^{-1}\left[F_{Mel}(f_l) + m\frac{F_{Mel}(f_h) - F_{Mel}(f_l)}{M+1} \right] \tag{9}$$

Where S is the number of sampling points, f_s is the sampling frequency, and f_l and f_h are the lowest and highest filter frequencies, respectively.

Because human auditory perception is logarithmic in amplitude, the Mel filter output is transformed logarithmically to simulate perceived loudness. The spectral energy at the output of the m^{th} Mel filter is calculated as:

$$S(i,m) = \text{In}\left[\sum_{k=0}^{N-1} E(i,k) H_m(k)\right], 0 \leq m \leq M \tag{10}$$

Finally, applying a Discrete Cosine Transform (DCT) to the logarithmic Mel energies produces the MFCC coefficients:

$$c(n) = \sum_{m-1}^{M} S(m)\cos(\frac{\pi n(m-0.5)}{M}), n = 0,1,2,...,K \tag{11}$$

DCT organises the signal energy in a set of low-order coefficients, which are appropriately related to the characteristics and significantly reduce the dimension of the data, which makes the subsequent classification tasks more robust and efficient.

2.2 Dynamic MFCC

Dynamic MFCCs enhance the traditional static MFCCs representation by incorporating temporal change information of the signal. Static MFCC coefficients are essentially equivalent to instantaneous acoustic characteristics, as they represent the spectral envelope of the signal within a single frame. However, they do not capture the temporal evolution of the signal's characteristics. In contrast, physiological signals such as voice and heart sound are characterised by strong time continuity, and the transition between states contains important diagnostic or classification information.

To explain this limitation, we only need to use first-order and second-order derivatives, Delta characteristics and Delta-Delta characteristics respectively. These quantify the speed and acceleration of spectral features and enhance the ability of the model to capture dynamic time characteristics.

The first-order difference (Delta feature) represents the rate of change of an MFCC coefficient over time, approximating its first-time derivative and indicating the envelope of transition between frames. Mathematically, this is expressed as:

$$\Delta c_t(n) = \frac{\sum_{\theta=1}^{\Theta} \theta[c_{t+\theta}(n) - c_{t-\theta}(n)]}{2\sum_{\theta=1}^{\Theta} \theta^2} \tag{12}$$

Here, $c_t(n)$ is the nth coefficient of MFCC at frame t, $\Delta c_t(n)$ is the first-order difference of $c_t(n)$ (or Delta coefficient), and K is the regression window size (which defines the temporal context by including adjacent frames). K is the time offset weight, which determines the impact of previous and current time frames.

The second-order difference (Delta-Delta characteristic) is the first-order rate of change of the Delta coefficient, which captures the acceleration of MFCC evolution and thus represents the second-order rate of change of MFCCs. The calculation of Delta features applies a differentiation operation:

$$\Delta^2 c_t(n) = \frac{\sum_{\theta=1}^{\Theta} \theta[c_{t+\theta}(n) - c_{t-\theta}(n)]}{2\sum_{\theta=1}^{\Theta} \theta^2} \tag{13}$$

Similarly, this encapsulates the second-order temporal dependence of the rate of change of spectral characteristics—a desirable feature to describe transient or periodic characteristics in heart sound signals.

2.3 Deep residual network (ResNet) 18

The ResNet was first introduced by He et al. in 2015, representing a major milestone in the design of deep neural networks [20]. A major innovation of ResNet is the use of residual connections, which mitigate the vanishing gradient problem and alleviate network degradation in deep networks during training. Thus, it enables the design of deeper and more stable architectures. Since its design, ResNet has achieved great success in many fields, including image classification, speech recognition and medical signal analysis. The basic/residual block is shown in **Figure 2**.

As shown in **Figure 2**, the residual block is formed by two paths: the residual mapping path composed of multiple layers (convolution, batch normalization, rectified linear unit, etc.), represented by $F(x)$, and the identity mapping path, in which the input x is directly passed to the output through the so-called skip connection. The output calculation of the residual block $h(x)$ is as follows:

$$y = F(x) + x \tag{14}$$

Here, $F(x)$ represents the nonlinear transformation of input x. The original input x is directly added to $F(x)$ via the skip connection, which allows direct feature propagation.

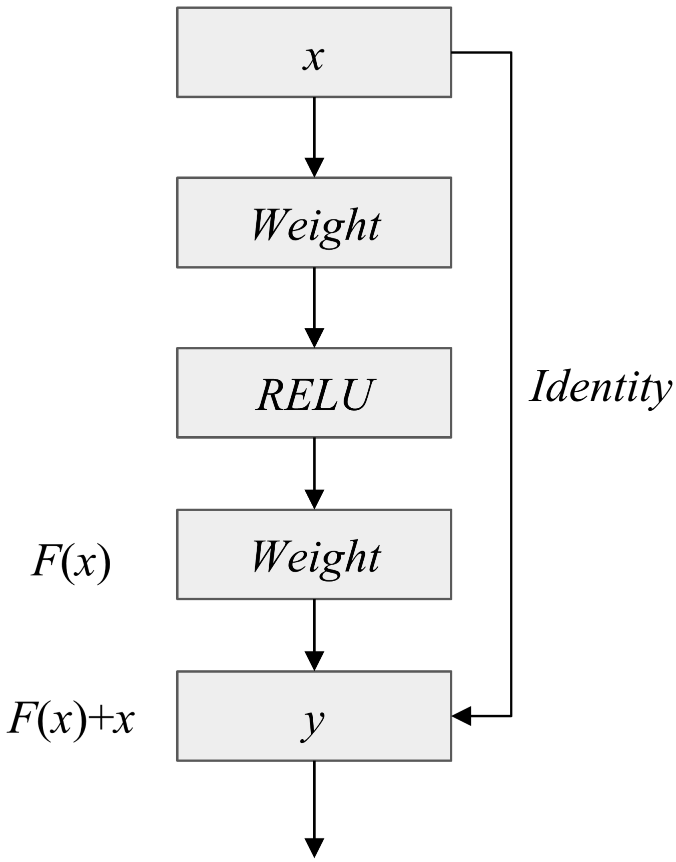


Figure 2. Residual module structure. RELU, rectified linear unit.

ResNet18 is one of the most prominent and simplest models in the ResNet family, comprising 18 layers. The architecture aims to trade some feature extraction capabilities to keep parameter counting and computing complexity relatively low, allowing it to be used for small-scale data sets (and/or embedded devices). The overall structure of ResNet18 is shown in **Figure 3**.

2.4 Attention mechanisms

The attention mechanism enables the deep learning model to focus on useful features while suppressing unrelated features and improving the performance of the model's representation learning and classification [21, 22]. Three types of attention modules are used in this work: frequency attention, Multi-Head Attention (MHA) and Coordinate Attention (CA).

2.4.1 Frequency attention mechanism

The frequency attention mechanism is designed for time–frequency representations such as MFCCs. It focuses on the frequency dimension to highlight discriminative spectral components.

Given an input feature map $X \in R^{C \times F \times T}$, temporal information is aggregated using global average pooling:

$$z_c(f) = \frac{1}{T}\sum_{t=0}^{T} X_c(f,t) \tag{15}$$

Then, two 1×1 convolutional layers with nonlinear activations generate frequency-wise attention weights:

$$a = \sigma(Conv_{1\times1}^{(2)}(\delta(Conv_{1\times1}^{1}(z)))) \tag{16}$$

Finally, the input features are reweighted by the attention map:

$$y_c(f,t) = a_c(f) \times x_c(f,t) \tag{17}$$

This mechanism adaptively enhances informative frequency bands while suppressing redundant information.

2.4.2 MHA mechanism

The MHA mechanism, introduced in the Transformer architecture, captures dependencies across multiple representational subspaces [23].

For input sequence X, the query, key, and value matrices are computed as:

$$Q = XW^Q, K = XW^K, V = XW^V \tag{18}$$

The attention output is then obtained as:

$$Attention(Q,K,V) = Soft\max(\frac{QK^T}{\sqrt{d_k}})V \tag{19}$$

Multiple attention heads are computed in parallel and concatenated:

$$MHA(X) = Concat(head1, ..., head_h)W^O \tag{20}$$

This structure enables the model to learn richer contextual relationships and improve generalization.

2.4.3 CA mechanism

The CA mechanism is a lightweight module that encodes both channel relationships and precise positional information by decomposing spatial attention into two independent directions—horizontal and vertical [24, 25].

Given an input feature map $X \in R^{C \times H \times W}$, CA first applies global average pooling along the height and width dimensions:

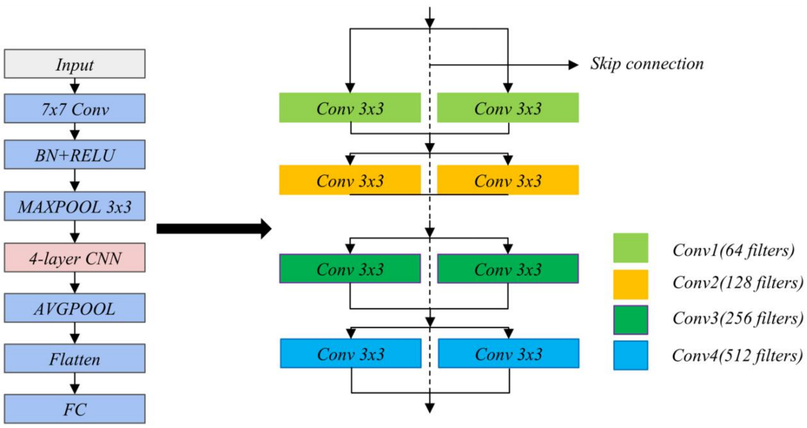


Figure 3. ResNet18 architecture. Conv, convolution; BN, batch normalization; RELU, rectified linear unit; MAXPOOL, max pooling; AVGPOOL, average pooling; FC, fully connected (layer).

$$z_c^h(h) = \frac{1}{W}\sum_{i=0}^{W-1} x_c(h,i) \tag{21}$$

$$z_c^w(w) = \frac{1}{H}\sum_{j=0}^{H-1} x_c(j,w) \tag{22}$$

The pooled features are concatenated and passed through a shared 1×1 convolution and rectified linear unit activation for feature transformation:

$$f = \delta(F_1(z^h, z^w)) \tag{23}$$

The transformed feature is then split into two tensors, corresponding to the horizontal and vertical directions.

$$f_h, f_w = split(f) \tag{24}$$

Two independent 1×1 convolutions with sigmoid activations generate directional attention weights:

$$a_h = \sigma(F_h(f_h)), a_w = \sigma(F_w(f_w)) \tag{25}$$

Finally, the input feature map is reweighted by the corresponding attention weights to produce the refined output:

$$Y(c,h,i) = X(c,h,i) \cdot a_h(c,h) \cdot a_w(c,i) \tag{26}$$

Using this separable coding method, CA captures long-range spatial dependencies while maintaining accurate location information and feature representation with lower computational complexity.

2.5 SVM

SVM is a supervised learning type model based on statistical learning theory. It develops an optimal hyperplane to maximise the separation between classes by establishing a large margin represented by a set of examples in a high-dimensional feature space. SVM learns a hyperplane, which provides powerful generalisation and classification performance based on structural risk minimisation.

Given a training set:

$$\{(x_i, y_i)\}_{i=1}^N, x_i \in \mathbb{R}^d, y_i \in \{-1, +1\} \tag{27}$$

SVM seeks the optimal parameters w and b by solving the following optimization problem:

$$\min_{w,b} \frac{1}{2}\| w \|^2 y_i(w^T x_i + b) \geq 1, i = 1,2,...,N \tag{28}$$

Using the Lagrange multiplier method, the dual form is expressed as:

$$\max \sum_{i=1}^N a_i - \frac{1}{2}\sum_{i=1}^N\sum_{j=1}^N a_i a_j y_i y_j (x_i^T x_j), \sum_{i=1}^N a_i y_i = 0, 0 \leq a_i \leq C \tag{29}$$

Where a_i are the Lagrange multipliers, and C is the penalty factor.

The final decision function is given by:

$$f(x) = sign(\sum_{i=1}^N a_i y_i K(x_i, x) + b) \tag{30}$$

This formula allows SVM to use any number of kernel functions, such as linear, polynomial and radial basis function kernels, to effectively handle both linear and nonlinear classification tasks.

Multi-Branch Fusion MFCC (MB-FusionMFCC) model

In this work, we propose a MB-FusionMFCC to utilise the temporal dynamics of heart sound signals. Compared with previous methods that directly concatenate and classify MFCCs along with their first- and second-order derivatives, we treat the three feature types—static MFCCs, Delta coefficients, and Delta-Delta coefficients—as independent input branches, enabling each branch to learn representations specific to its characteristics. The model introduces a cross-branch attention model to improve the interaction and integration of feature types. The overall architecture is presented in **Figure 4**, and the detailed structure of each branch is shown in **Figure 5**.

The convolutional feature extraction module is composed of a series of convolutional, batch normalisation, gaussian error linear unit, and pooling layers, enabling the model to learn time-frequency hierarchies.

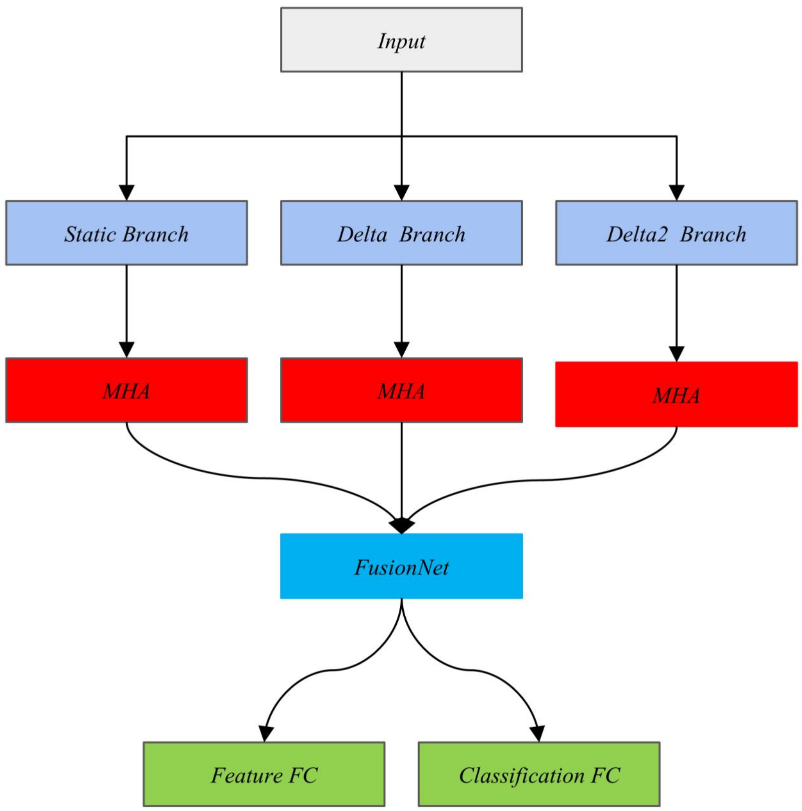


Figure 4. MB-FusionMFCC structure. Each branch shares the same architecture, which consists of three key modules: (1) a convolutional feature extraction module, (2) a frequency attention module, and (3) a global feature mapping module. MHA, Multi-Head Attention; FC, fully connected (layer); MB-FusionMFCC, multi-branch fusion mel-frequency cepstral coefficient.

The frequency attention module performs global pooling across time dimensions to learn the adaptive weighting of the frequency band, allowing the model to emphasize more discriminative frequency components for classification.

The global feature mapping module executes the global average pool and uses a fully connected projection to compress the two-dimensional feature map into a fixed-length vector to form a compact identification feature embedding.

The outputs of all branches are then integrated via a cross-branch attention module, which captures interactions among different feature representations by generating an attention-weighted fusion. Finally, the representation of cross-fusion is connected, then exited through the fusion layer, and then normalised through the weight mapping process.

The final result is to classify the final fusion feature vector into the heart sound category. It is also very reasonable to suspect that this advanced fusion representation may also exist in other tasks, such as diagnostic tasks or feature visualisation.

2.6 CA-ResNet18

This study uses CA-ResNet18, a CNN augmented with a CA fusion module, to extract the characteristics of classification tasks with dynamic MFCC feature images. The model is based on the structure of the standard ResNet18. By including the CA module, the model enhances the sensitivity to informative regions and the awareness of positional information within the spatial feature maps for classification.

Compared with the standard channel attention mechanism, the CA module pays global context attention to the height and width location information, while maintaining the location information of the spatial feature map. More importantly, the CA module decomposes spatial information through separate horizontal and vertical pooling on the 2D feature map, then globally encodes and recalibrates the features. This provides spatially-aware attention weights that enhance task-relevant regions.

In terms of architectural design, CA-ResNet18 uses the original definition of the ResNet18 residual block, but includes the CA module in the third (layer 3) and fourth (layer 4) stages of the model. Each residual unit is designed to first extract local features with convolution and batch normalization layers, and then use the CA module to provide joint channel attention and spatial attention for map features, so as to improve discrimination ability and robustness. **Figure 6** shows the structural representation of the remaining units containing the CA module.

The final output of the network is projected into the feature vector, which is passed to the classifier to predict class labels. This allows CA-ResNet18 to focus on the area most relevant to the heart sound pattern in the dynamic MFCC image, so as to realise the multi-dimensional representation of time and frequency characteristics to improve the classification performance.

3 EXPERIMENTAL METHODS AND EVALUATION METRICS

3.1 Experimental setup

The experimental simulation platform is configured using PyCharm Professional 2024.2.4, and the network is trained and evaluated on the PyTorch framework. The hardware includes NVIDIA RTX 4080 SUPER GPU, AMD Ryzen 9 9950X 16-core CPU, running Windows 11 and Python 3.9 as the programming language.

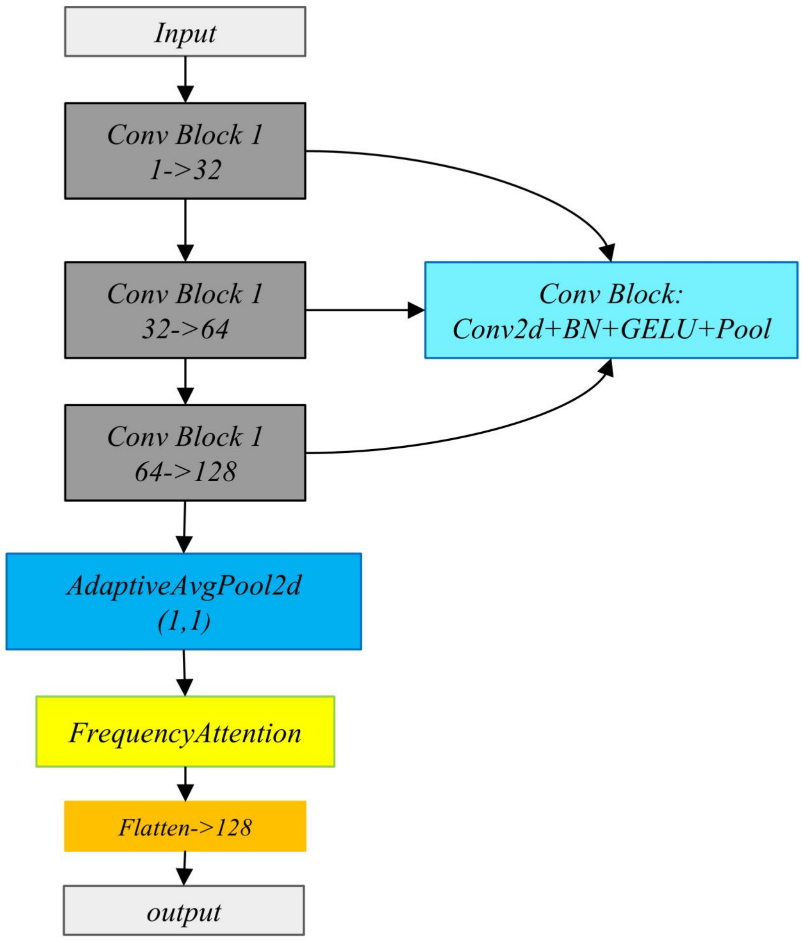


Figure 5. Branch structure. Conv, convolution (layer); BN, batch normalization; GELU, gaussian error linear unit; Pool, pooling (layer); AdaptiveAvgPool2d, adaptive average pooling 2D.

The model performance evaluation is carried out using five-fold cross-validation, which divides the dataset into five folds of equal size to maintain the same proportion of normal and abnormal heart sounds. In each iteration, one fold is regarded as a test set, and the remaining four folds are regarded as a training set to fully evaluate the generalisation ability of the model.

3.2 Evaluation metrics

The proposed method evaluation is carried out using the standard classification indicators of accuracy (*Acc*), sensitivity (*Se*), specificity (*Sp*), precision (*Pre*) and F1 score (*F1*). The definitions of these indicators are as follows:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{31}$$

$$Se = \frac{TP}{TP + FN} \tag{32}$$

$$Sp = \frac{TN}{TN + FP} \tag{33}$$

$$Pre = \frac{TP}{TP + FP} \tag{34}$$

$$Recall = Se \tag{35}$$

$$F1 = 2 \times \frac{Pre \times Re\ call}{Pre + Re\ call} \tag{36}$$

Here, TP, TN, FP and FN are used to indicate true positive, true negative, false positive and false negative respectively. These indicators together evaluate the accuracy of classification and the performance of the model in classifying abnormal heart sounds.

3.3 Dataset

The experiment was conducted on the public dataset from the 2016 PhysioNet/CinC Challenge, which comprises six subsets with a total of 3,240 heart sound recordings. Of these, 2,575 recordings are considered normal heart sounds, and 665 are considered abnormal heart sounds [26]. The length of each heart recording ranges from 5 seconds to 122 seconds, and the sampling rate is 2 kHz. For each subset, the distribution of heart sound recordings is shown in **Table 1**.

3.4 Data preprocessing

Data preprocessing in this study consists of three main steps: filtering, segmentation, and normalization.

First of all, the data shows that the acquisition noise is mainly related to the stethoscope rubbing the subject's skin at the beginning and end of recording. Therefore, the first second of each recording at both the beginning and the end has been deleted. The heart sound signals are segmented into three-second fragments, and any fragments shorter than three seconds are discarded. The signal is band-pass filtered using a 20-400 Hz fifth-order Butterworth filter to minimise high-frequency noise while maintaining the main frequency component of the heart sound signal.

Secondly, the dataset shows a strong imbalance, and the normal sample is much richer than the abnormal sample. Therefore, a targeted data balance strategy has been implemented. For normal samples, a fragment was obtained from each record, and for abnormal samples, four fragments were obtained to try to balance the representativeness of the minority. The datasets of training, verification and testing are divided at the patient level to avoid the overrepresentation of any individual sample and eliminate the overlap between patients.

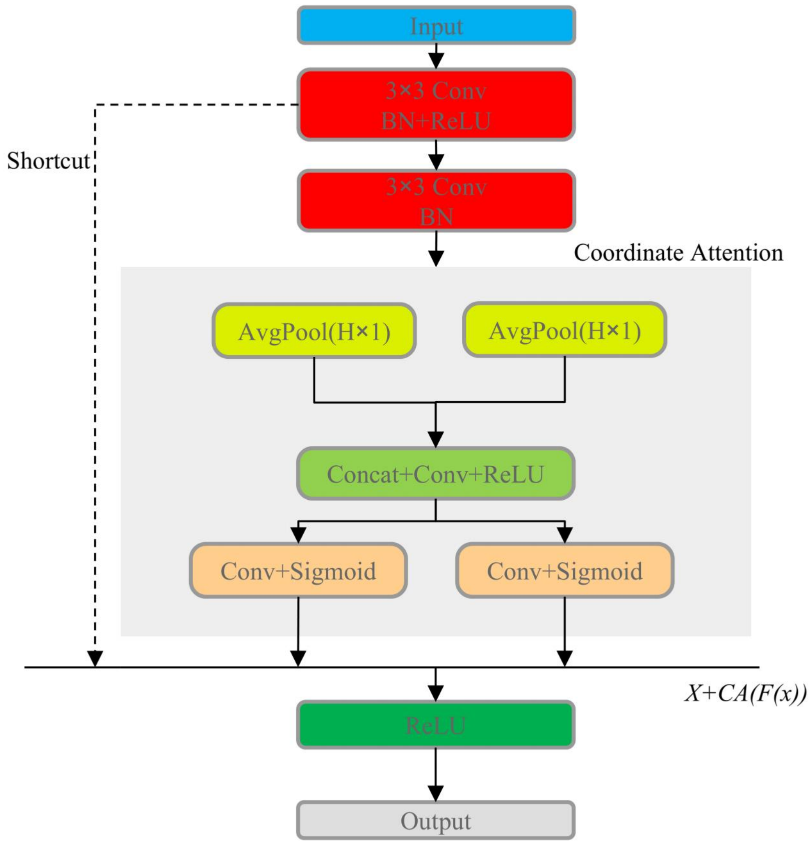


Figure 6. Residual unit structure after CA module embedding. CA, Coordinate Attention; Conv, convolution (layer); BN, batch normalization; ReLU, rectified linear unit; AvgPool, average pooling; Concat, concatenation (operation).

Table 1. Statistics for each subset of the PhysioNet challenge 2016 dataset

Dataset	Type			Duration (seconds)		
	Normal	Abnormal	Total	Minimum	Maximum	Average
Training-a	117	292	409	9	36	33
Training-b	368	104	490	5	8	8
Training-c	7	24	31	10	122	49
Training-d	27	28	55	7	49	15
Training-e	1,958	183	2,141	8	102	23
Training-f	80	34	114	29	60	33
Total	2,575	665	3,240	5	122	22

Finally, all fragments of the heart sound are normalised in range to unify the numerical scale across samples to prevent gradient instability due to large amplitude denaturation, which helps to provide the convergence of the model and ultimately improve the classification performance.

3.5 MFCC static and dynamic feature extraction

Once the data is preprocessed, the MFCC characteristics are extracted. Each three-second segment is sampled at 2,000 Hz, then divided into overlapping frames of 256 samples with a frame shift of 64 samples, generating about 90 time frames. Each frame is converted to the frequency domain by applying Fast Fourier Transform, and then the Mel filter bank is used to compute the logarithmic power spectrum. The resulting logarithmic power spectrum runs through DCT, and the 13-dimensional static MFCC coefficient is obtained.

From the static MFCCs, the first-order and second-order difference characteristics are calculated to characterise the time dynamics of spectral changes. These estimates are connected with static coefficients to produce a three-dimensional dynamic MFCC feature vector with 39 characteristics. In addition, all features have been normalised on a scale to ensure consistent statistical scaling across samples and produce a unified input format that can be visualised to develop training methods.

3.6 Extraction of static and dynamic MFCC feature images

After obtaining the 39-dimensional dynamic MFCC feature, all the feature vectors corresponding to the given time frame are connected in series according to the time order corresponding to x, and a two-dimensional matrix of size (39, 90) is generated, of which 39 corresponds to the frequency dimension and 90 is the time frame number, which acts as a heart sound signal. The two-dimensional time frequency is presented, and the time evolution of the spectral energy distribution is presented.

The matrix is linearly normalised to eliminate the amplitude changes of all samples in the pseudo-colour image. Then the normalised and numerically grounded components are quantitatively mapped to pixels in order to extract the two-dimensional demonstration of pseudo-colour and visually distinguish the characteristics. Compared with static MFCC images, dynamic MFCC images incorporate first- and second-order differential information into the original spectral structure, thereby capturing both instantaneous and transitional acoustic characteristics. This enriched representation provides deeper temporal context and a more discriminative feature space for deep neural network learning, the example of dynamic MFCC feature images is presented in **Figure 7**.

4 EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Heart sound classification based on single features

In classification tasks based on MFCC and image features, CNNs and their derived architectures remain the most representative mainstream frameworks. To further evaluate the performance of the proposed models, **Tables 2** and **3** present the

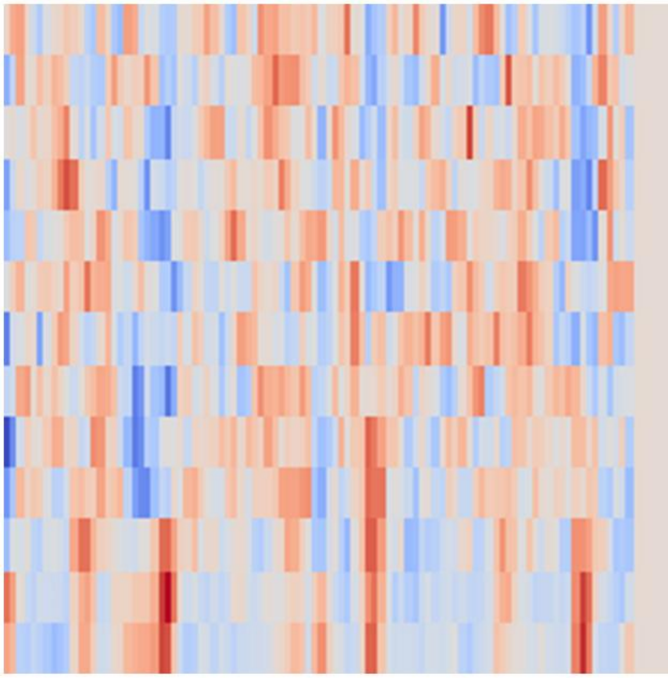


Figure 7. Examples of dynamic MFCC feature images. MFCC, mel-frequency cepstral coefficient.

Table 2. Performance comparison of different CNN models on dynamic MFCC features

Model	F1	Acc	Pre	Se	Sp
ResNet18	0.8339	0.8439	0.8440	0.8240	0.8620
ResNet34	0.8459	0.8557	0.8595	0.8326	0.8767
ResNet50	0.8455	0.8501	0.8288	0.8629	0.8386
Inception_v3	0.8398	0.8378	0.7933	0.8909	0.7896
MobileNet_v2	0.8310	0.8475	0.8785	0.7883	0.9012
MB-FusionMFCC	0.8776	0.8768	0.8317	0.9287	0.8397

Note: Acc, accuracy; Pre, precision; Se, sensitivity; Sp, specificity; MB-FusionMFCC, multi-branch fusion mel-frequency cepstral coefficient; CNN, convolutional neural network; MFCC, mel-frequency cepstral coefficient.

Table 3. Performance comparison of different CNN models on dynamic MFCC image features

Model	F1	Acc	Pre	Se	Sp
ResNet18	0.8567	0.8552	0.8090	0.9104	0.8053
ResNet34	0.8425	0.8501	0.8416	0.8434	0.8562
ResNet50	0.8365	0.8352	0.7917	0.8866	0.7886
MobileNetV2	0.8357	0.8362	0.7990	0.8758	0.8004
SE-ResNet18	0.8464	0.8511	0.8306	0.8629	0.8405
CA- ResNet18	0.8708	0.8727	0.8410	0.9028	0.8454

Note: Acc, accuracy; Pre, precision; Se, sensitivity; Sp, specificity; CNN, convolutional neural network; MFCC, mel-frequency cepstral coefficient; SE-ResNet18, ResNet18 with Squeeze-and-Excitation module; CA-ResNet18, ResNet18 with Coordinate Attention module.

Table 4. Performance comparison of different CNN-SVM models on dynamic MFCC features

Model	F1	Acc	Pre	Se	Sp
ResNet18-SVM	0.9164	0.9177	0.8902	0.9442	0.8934
ResNet34-SVM	0.9282	0.9299	0.9079	0.9495	0.9118
ResNet50-SVM	0.9020	0.9012	0.8569	0.9524	0.8545
Inception_v3-SVM	0.9398	0.9418	0.9295	0.9504	0.9340
MobileNet_v2-SVM	0.8982	0.8981	0.8588	0.9415	0.8583
MB-FusionMFCC-SVM	0.9626	0.9641	0.9594	0.9658	0.9625

Note: Acc, accuracy; Pre, precision; Se, sensitivity; Sp, specificity; CNN, convolutional neural network; MFCC, mel-frequency cepstral coefficient; SVM, support vector machine; MB-FusionMFCC, multi-branch fusion mel-frequency cepstral coefficient.

Table 5. Performance comparison of different CNN-SVM models on dynamic MFCC image features

Model	F1	Acc	Pre	Se	Sp
ResNet18-SVM	0.9604	0.9619	0.9522	0.9688	0.9555
ResNet34-SVM	0.9112	0.9116	0.8761	0.9496	0.8772
ResNet50-SVM	0.8910	0.8908	0.8513	0.9350	0.8506
MobileNet_v2-SVM	0.9468	0.9484	0.9321	0.9621	0.9360
SE-ResNet18-SVM	0.9564	0.9577	0.9431	0.9702	0.9464
CA- ResNet18-SVM	0.9643	0.9656	0.9564	0.9724	0.9594

Note: Acc, accuracy; Pre, precision; Se, sensitivity; Sp, specificity; CNN, convolutional neural network; MFCC, mel-frequency cepstral coefficient; SVM, support vector machine; SE-ResNet18, ResNet18 with Squeeze-and-Excitation module; CA-ResNet18, ResNet18 with Coordinate Attention module.

classification results of the MB-FusionMFCC and CA-ResNet18 models compared with several typical CNN models and their improved variants.

In the classification tasks using dynamic MFCC features and dynamic MFCC image features, the experimental results show that CNNs and their improved variants can effectively extract discriminative representations, though performance differences exist across architectures. For dynamic MFCC features, the MB-FusionMFCC model outperforms the traditional ResNet series, Inception_v3, and MobileNet_v2 in both F1-score and accuracy, indicating that its multi-branch fusion structure enhances the representation capability of MFCC features.

For dynamic MFCC image features, CA-ResNet18 achieves better performance than the original ResNet18 and other attention mechanisms (e.g., Squeeze-and-Excitation, Efficient Channel Attention), demonstrating that the CA mechanism helps capture both local and global dependencies in time–frequency representations.

Next, we introduce an SVM classifier on top of the existing models, where deep learning is used for feature extraction, and SVM serves as a lightweight classifier for final decision-making. The results are shown in **Tables 4** and **5**.

Table 6. Classification performance comparison between fusion feature and single feature models

Model	Feature type	F1	Acc	Pre	Se	Sp
MB-Fusion MFCC-SVM	Dynamic MFCC Features	0.9626	0.9641	0.9594	0.9658	0.9625
CA-ResNet18-SVM	Dynamic MFCC Image	0.9643	0.9656	0.9564	0.9724	0.9594
Fusion-SVM (this paper)	Dynamic MFCC Fusion	0.9670	0.9682	0.9591	0.9751	0.9619

Note: Acc, accuracy; Pre, precision; Se, sensitivity; Sp, specificity; MFCC, mel-frequency cepstral coefficient; SVM, support vector machine; MB-Fusion MFCC-SVM, multi-branch fusion MFCC model with SVM classifier; CA-ResNet18-SVM, Coordinate Attention ResNet18 with SVM classifier; Fusion-SVM, fusion model with SVM classifier.

Table 7. Classification performance comparison between proposed and existing methods

Model	Feature type	Acc	Se	Sp
SVM [4]	Gamma Spectrum Features	0.9336	0.9752	0.7560
CAFusionNet [15]	MFCC+Mel-spectrogram	0.9665	0.9667	0.9665
MDN-MARNN [10]	Raw Heart Sound Data	0.9541	0.9400	0.9681
1D-CNN+WST [11]	WST+MFCC+STFT	0.9651	0.9660	0.9660
CTENN [12]	Raw Heart Sound Data	0.9640	0.9287	0.9745
MobileNet+Dense121 [16]	Spectrogram	0.9567	0.9313	0.9821
This Study	Dynamic MFCC+Image Features	0.9682	0.9751	0.9619

Note: Acc, accuracy; Se, sensitivity; Sp, specificity; SVM, support vector machine; MFCC, mel-frequency cepstral coefficient; CAFusionNet, Channel Attention fusion network; MDN-MARNN, Multi-scale DenseNet–Multi-head Attention Recurrent Neural Network; 1D-CNN, 1D-convolutional neural network; WST, wavelet scattering transform; CTENN, CNN-Transformer End-to-end Neural Network; STFT, short-time fourier transform.

As shown in **Tables 4** and **5**, after introducing the SVM classifier, both MB-FusionMFCC and CA-ResNet18 models exhibit a significant improvement in classification performance, with F1-scores and accuracies exceeding 0.96. This demonstrates that deep neural networks are highly effective in extracting high-level representations, while the lightweight SVM classifier enhances the discriminative capability and generalization performance of the overall system.

4.2 Heart sound classification based on dynamic MFCC and dynamic MFCC image feature fusion

In the preceding sections, independent classification experiments were conducted using MB-FusionMFCC-SVM and CA-ResNet18-SVM for dynamic MFCC features and dynamic MFCC image features, respectively. Both models show high classification accuracy, which shows that they have strong feature representation and discrimination in both the time domain and the time-frequency domain.

Heart sound samples are used to extract advanced feature vectors through MB-FusionMFCC and CA-ResNet18 models. Each feature set is normalised before being connected together, resulting in a fusion feature representation. Fusion feature vectors are accessed by the SVM classifier for final classification.

Table 6 summarises the classification performance of the fusion feature model and the single feature model, and shows the benefits of multimodal feature fusion for heart sound classification.

It can be seen from the survey results that when these features are input into Fusion-SVM, the combination of features extracted from the two models exceeds the single-modal use case. In order to check the validity of the proposed feature combination method, it was compared with the existing models, which represent the cross-section of the cardiac sound classification model in the current literature. **Table 7** summarises the performance of the proposed model and the existing model trained on the common public data set (PhysioNet 2016 public dataset).

As shown in **Table 7**, our cardiac sound classification method uses dynamic MFCC characteristics and dynamic MFCC images to show very high accuracy, sensitivity and specificity. The accuracy of the proposed method reaches 0.9682, which is comparable to other recent methods (for example, CAFusionNet, Acc=0.9665), demonstrating its competitiveness in modern heart sound classification. Reference 4 was among the first to introduce gamma spectral features using the SVM classifier, and achieves an accuracy of 0.9336. The contribution of reference 4 lies in the design of features. However, its low specificity (0.7560) may be due to class imbalance. The models in MDN-MARNN and CTENN takes fragments of the original heart sound, which allows them to extract time characteristics globally [10, 12]. However, potentially due to noise and dataset complexity, their reported sensitivity or accuracy is low. The works in 1D-CNN+WST and MobileNet+Dense121 also use convolutional spectral features, achieving high accuracy and specificity, but their sensitivity is slightly lower than that of our method [11, 16]. Finally, CAFusionNet integrates MFCC features and Mel spectral map features to achieve accuracy (0.9665) and sensitivity (0.9667) [15]. While its sensitivity is comparable, its accuracy is slightly lower than our method's (0.9682). The authors note a limitation in the general detection of abnormal heart sounds.

In this study, we propose a new hybrid framework, which uses deep networks to extract advanced features, which are then classified by an SVM. The main advantage lies in combining the representational strength of deep learning with the robustness of an SVM classifier, even in high-dimensional spaces with limited sample sizes.

Contrary to most studies using spectral images, we use dynamic MFCC feature images instead. For example, static MFCC image characteristics alone will lead to poor accuracy, but dynamic MFCC images include a detailed representation of short-term frequency distribution, which can better represent time evolution and improve sensitivity to small changes in heart sounds.

Finally, we also use medium-term fusion. Each feature type extracts its characteristics separately before classification, so it is confirmed that the fusion feature will be better than a single feature model in terms of accuracy, sensitivity and specificity. These results are expected and demonstrate the complementary nature of the time-frequency characteristics of the image (or feature) and the effectiveness of the medium-term fusion process.

In summary, the results show that the proposed methods, including dynamic MFCC feature images and mid-level feature fusion, effectively capture both temporal and time-frequency characteristics of cardiac sounds. The introduction of the SVM classifier further enhances the discriminative capability and generalization of the models. Overall, these findings demonstrate the complementary nature of the fused features and set the stage for the subsequent conclusion, which provides a comprehensive summary of the study's main contributions and performance.

5 CONCLUSION

In this study, we proposed a cardiac sound classification method based on the mid-fusion of dynamic MFCC features and dynamic MFCC images. This method uses deep CNNs to extract features, yielding an advanced representation of time-frequency features, which can effectively detect the slight difference in heart sound signals between time and frequency. For classification, an SVM replaces the traditional fully connected layer, thus leveraging the robustness of a machine learning classifier within the deep learning framework. For example, SVM classifiers can improve robustness when learning from limited samples and reduce the overfitting of high-dimensional feature sizes.

Experimental results demonstrate that the proposed Fusion-SVM model achieves an F1-score of 0.9670, accuracy of 0.9682, sensitivity of 0.9751, and specificity of 0.9619 on the PhysioNet 2016 dataset. These results outperform single-feature models and most existing methods, confirming the effectiveness of mid-level feature fusion and the integration of deep feature representation with SVM classification.

Although we have achieved commendable results, several limitations exist. First of all, mid-term fusion is conceptually effective for integrating multi-source features, the increased dimensionality leads to higher computational overhead during training and testing. Therefore, future work could focus on selecting features or reducing dimensions to delete redundant information. Secondly, all experiments were conducted on the PhysioNet 2016 dataset; further validation on clinical cardiac sound data collected from multiple locations and conditions is needed to assess generalisation. Finally, this study did not explore all complementary acoustic features beyond dynamic MFCCs and images. Future work may integrate multimodal features to improve the robustness and adaptability of the model.

DECLARATIONS

Author contributions

Shoucheng Chen constructed and trained the model, and drafted the manuscript. Ke Wang and Wenjing Du contributed to data analysis and statistics, and revised the manuscript. Rongguo Yan provided guidance on experimental methods and critically revised the manuscript.

Funding

This research received no external funding.

Data availability

The dataset used in this study is publicly available from the PhysioNet/CinC Challenge 2016 database.

Ethics approval and consent to participate

Not applicable. The study used a publicly available dataset.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

Not applicable.

REFERENCES

[1] Chinese Society of Cardiology. Summary of China cardiovascular health and disease report 2024. Chin Circ J. 2025 Jul 9;40(6):521-559. https://doi.org/10.3969/j.issn.1000-3614.2025.06.001

[2] Li J, Ke L, Du Q. Classification of heart sounds based on the wavelet fractal and twin support vector machine. Entropy (Basel). 2019 May 6;21(5):472. https://doi.org/10.3390/e21050472

[3] Xu W, Yu K, Ye J, Li H, Chen J, Yin F, et al. Automatic pediatric congenital heart disease classification based on heart sound sig-

nal. Artif Intell Med. 2022 Apr;126:102257. https://doi.org/10.1016/j.artmed.2022.102257
[4] Taneja K, Arora V, Verma K. Heart sound classification method using gammatonegram and SVM. Multimed Tools Appl. 2025 Jun 1;84(21):23987-24023. https://doi.org/10.1007/s11042-024-19984-1
[5] Rubin J, Abreu R, Ganguli A, Nelaturi S, Matei I, Sricharan K. Classifying heart sound recordings using deep convolutional neural networks and mel-frequency cepstral coefficients. 2016 Computing in Cardiology Conference (CinC); 2016 Sep 11-14; Vancouver (BC). IEEE; 2016. p. 813-816.
[6] Deng M, Meng T, Cao J, Wang S, Zhang J, Fan H. Heart sound classification based on improved MFCC features and convolutional recurrent neural networks. Neural Netw. 2020 Oct;130:22-32. https://doi.org/10.1016/j.neunet.2020.06.015
[7] Ismail S, Ismail B, Siddiqi I, Akram U. PCG classification through spectrogram using transfer learning. Biomed Signal Process Control. 2023 Jan 1;79:104075. https://doi.org/10.1016/j.bspc.2022.104075
[8] Xiao B, Xu Y, Bi X, Zhang J, Ma X. Heart sounds classification using a novel 1-D convolutional neural network with extremely low parameter consumption. Neurocomputing. 2020 Jun 7;392:153-159. https://doi.org/10.1016/j.neucom.2018.09.101
[9] Oh SL, Jahmunah V, Ooi CP, Tan RS, Ciaccio EJ, Yamakawa T, et al. Classification of heart sound signals using a novel deep WaveNet model. Comput Methods Programs Biomed. 2020 Nov;196:105604. https://doi.org/10.1016/j.cmpb.2020.105604
[10] Li S, Sun J, Yang H, Pan J, Guo T, Wang W. Interpretable End-to-End heart sound classification. Measurement. 2024 Sept 30;237:115113. https://doi.org/10.1016/j.measurement.2024.115113
[11] Patwa A, Rahman MMU, Al-Naffouri TY. Heart murmur and abnormal PCG detection via wavelet scattering transform and 1D-CNN. IEEE Sens J. 2025;25(7):12430-12443. https://doi.org/10.1109/JSEN.2025.3541320
[12] Cheng J, Sun K. Heart sound classification network based on convolution and transformer. Sensors (Basel). 2023 Sept 29;23(19):8168. https://doi.org/10.3390/s23198168
[13] Abbas S, Ojo S, Al Hejaili A, Sampedro GA, Almadhor A, Zaidi MM, et al. Artificial intelligence framework for heart disease classification from audio signals. Sci Rep. 2024 Feb 7;14(1):3123. https://doi.org/10.1038/s41598-024-53778-7
[14] Lee JA, Kwak KC. Heart sound classification using wavelet analysis approaches and ensemble of deep learning models. Appl Sci. 2023;13(21):11942. https://doi.org/10.3390/app132111942
[15] Li M, He Z, Wang H. Heart sound classification based on multi-scale feature fusion and channel attention module. Bioengineering (Basel). 2025 Mar 14;12(3):290. https://doi.org/10.3390/bioengineering12030290
[16] Khan SUR, Khan Z. Detection of abnormal cardiac rhythms using feature fusion technique with heart sound spectrograms. J Bionic Eng. 2025 Jul 1;22(4):2030-2049. https://doi.org/10.1007/s42235-025-00714-8
[17] Huai X, Jiang L, Wang C, Chen P, Li H. Heart sound classification based on convolutional neural network with convolutional block attention module. Front Physiol. 2025 Jun 5;16:1596150. https://doi.org/10.3389/fphys.2025.1596150
[18] Althaph B, Challa NP. Explainable attention-based deep learning for classification and interpretation of heart murmurs using phonocardiograms. Sci Rep. 2025 Oct 30;15(1):37991. https://doi.org/10.1038/s41598-025-21971-x
[19] Fernandes J, Teixeira F, Guedes V, Junior A, Teixeira JP. Harmonic to noise ratio measurement - selection of window and length. Procedia Computer Science. 2018 Jan 1;138:280-285. https://doi.org/10.1016/j.procs.2018.10.040
[20] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016 Jun 27-30; Las Vegas (NV). IEEE; 2016. p. 770-778.
[21] Ranipa K, Zhu WP, Swamy MNS. A novel feature-level fusion scheme with multimodal attention CNN for heart sound classification. Comput Meth Prog Bio. 2024 May 1;248:108122. https://doi.org/10.1016/j.cmpb.2024.108122
[22] Wu T, Huang Z, Li S, Zhao Q, Pan F. Heart murmur quality detection using deep neural networks with attention mechanism. Appl Sci. 2024;14(15):6825. https://doi.org/10.3390/app14156825
[23] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Advances in Neural Information Processing Systems 30 (NIPS 2017); 2017 Dec 4-9; Long Beach (CA). Neural Information Processing Systems Foundation, Inc. (NeurIPS); 2017.
[24] Sun L, Lei Y, Zhang Z, Tang Y, Wang J, Ye L, et al. Multi-task coordinate attention gating network for speech emotion recognition under noisy circumstances. Biomed Signal Proces. 2025 Sep 1;107:107811. https://doi.org/10.1016/j.bspc.2025.107811
[25] Wu J, Li X, Li T, Meng F, Kong Y, Yang G, et al. CSLNSpeech: Solving the extended speech separation problem with the help of Chinese sign language. Speech Commun. 2024 Nov 1;165:103131. https://doi.org/10.1016/j.specom.2024.103131
[26] Liu C, Springer D, Li Q, Moody B, Juan RA, Chorro FJ, et al. An open access database for the evaluation of heart sound algorithms. Physiol Meas. 2016 Dec;37(12):2181-2213. https://doi.org/10.1088/0967-3334/37/12/2181

Progress in
Medical Devices

Zentime

REVIEW ARTICLE

Research progress on hemostatic techniques for combat trauma

Xinying Shi[1], Yuan Yao[2], Haipo Cui[1]

[1]Shanghai Institute for Minimally Invasive Therapy, University of Shanghai for Science and Technology, Shanghai 200093, China.
[2]Shanghai Songyu Medical Devices Co., Ltd., Shanghai 200050, China.

Corresponding author: Haipo Cui.

Address correspondence to: Haipo Cui, Shanghai Institute for Minimally Invasive Therapy, University of Shanghai for Science and Technology, No. 516 Jungong Road, Yangpu District, Shanghai 200093, China.
E-mail: h_b_cui@163.com.

Received May 13, 2025; Accepted November 21, 2025; Published March 31, 2026

DOI: 10.61189/368729kpldnv

Abstract

Combat trauma hemostasis techniques are crucial for modern military medicine in addressing the challenges posed by high-energy destructive weapons. As reported, the incidence of vascular injuries during the Vietnam War was approximately 2%, while that during the Iraq War ranged from 4.4% to 8.2%. Meanwhile, massive hemorrhage caused by various types of vascular injuries is the primary factor leading to acute death among potentially survivable casualties during wartime. This paper provides a detailed exploration of the research progress in combat trauma hemostasis, with a focus on analyzing vascular injuries and uncontrolled bleeding caused by high-energy destructive weapons in modern warfare. Starting from the pathophysiological characteristics of combat trauma, it introduces the four-level priority treatment system established by North Atlantic Treaty Organization and Committee on Tactical Combat Casualty Care, and emphasizes the importance of the "Golden 1-Hour" principle in improving casualty survival rates. The physiological mechanisms of blood coagulation are outlined, followed by an in-depth analysis of both traditional and novel hemostatic techniques, including tourniquets, hemostatic dressings, and auxiliary hemostatic materials. Finally, the challenges faced by combat trauma hemostasis technologies are discussed, along with future development directions.

Keywords: Combat trauma, Hemostasis techniques, Physiological mechanisms, Hemostatic materials

1 INTRODUCTION

In the injury spectrum of modern warfare, explosive fragments and gunshot wounds account for over 75% of battlefield casualties, where uncontrollable hemorrhage is the primary pathological factor leading to on-site fatalities among casualties [1]. Notably, with the continuous iteration and upgrading of high-energy destructive weapons, the incidence of vascular injuries has shown a significant upward trend. During the Vietnam War, the occurrence rate of vascular trauma was merely 2% [2, 3]. During the Iraq War, this figure rose to 4.4%-8.2% [4, 5]. More recent analyses from conflicts in Afghanistan and Syria have suggested that vascular injury rates can reach 9-12% in conventional combat scenarios [6]. Research on combat trauma treatment has confirmed that fatal massive hemorrhages caused by various types of vascular injuries are the primary cause of acute death among potentially survivable battlefield casualties, with their pathological significance far exceeding that of other trauma types [7].

The evolution of modern warfare has further intensified the complexity of combat casualty care. Under the three-dimensional combat mode supported by sixth-generation communication technology and the application of new equipment such as high-explosive munitions and thermobaric weapons, casualties commonly exhibit the "four highs" characteristics—namely, a high proportion of kinetic energy injuries (>68%), a high incidence of combined injuries (41%), a high degree of tissue destruction, and a high risk of secondary hemorrhage [7, 8]. Taking a 70 kg healthy adult male as an example, the circulating blood volume accounts for approximately 7%-8% of body weight (about 4,900 ml). When acute blood loss exceeds 30% of the total blood volume (i.e., 1,470 ml), it can lead to decompensated shock. Among battlefield fatalities, 56% are directly attributed to uncontrolled bleeding, with this proportion rising to as high as 83% in preventable prehospital deaths [9]. Clinical big data analysis reveals that among fatal complications such as multiple organ dysfunction syndrome and disseminated intravascular coagulation following combat trauma, 61.2% of cases involve uncontrolled progressive hemorrhage, with over 18% of these deaths potentially preventable through early and effective hemostatic intervention [10].

Given the unique pathophysiological characteristics of combat injuries, domestic vascular surgery experts, based on the analysis of outcomes from the management of vascular combat trauma during years of military operations and training, propose that the treatment of vascular combat trauma should adhere to the "Three Rapid" principles: "rapid diagnosis, rapid hemostasis, and rapid revascularization" [3]. The Committee on Tactical Combat Casualty Care of North Atlantic Treaty Organization has established a four-tier priority management system: Immediate, Delayed, Minimal, and Expectant [11]. These principles and systems operate within a framework constrained by three major contradictions in combat casualty care: 1) the contradiction between the probability of mass casualties and limited medical resources; 2) the contradiction between complex environments (such as high-altitude hypoxia and high humidity in maritime areas) and the effectiveness of hemostatic materials; 3) the contradiction between multi-echelon evacuation (including three levels of care: tactical zone, operational zone, and strategic zone) and the continuity of hemostatic measures. Under this challenging framework, the "Golden Hour Rule" emphasizes that for every 10-minute reduction in the time window from injury to definitive hemostasis, the survival rate of casualties increases by 7.3% [12]. This stringent timeliness requirement arises directly from the need to overcome these contradictions. Therefore, the development of a dual-mode hemostatic technology system that combines "rapid response in frontline first aid" with "compatibility and scalability for rear surgical procedures" has become a strategic priority in 21st-century military medical research.

2 OVERVIEW OF THE PHYSIOLOGICAL MECHANISMS OF HEMOSTASIS

Hemostasis on the battlefield involves a combination of mechanical occlusion, biochemical coagulation, and thermal method, all ultimately relying on the body's intrinsic coagulation cascade. Vascular occlusion primarily relies on physical methods, with compression being the main hemostatic approach at body junctions [13]. For instance, after bleeding occurs, rapid hemostasis is achieved by using fingers to compress the superficial area of an artery overlying a bony surface (proximal end), thereby blocking blood flow. This method is simple to perform and suitable for hemostasis in areas such as the head, neck, shoulders, and limbs, but it is difficult to maintain for extended periods. Other tools like tourniquets and hemostatic forceps can also be employed to swiftly cut off blood flow at the bleeding site for rapid hemostasis. Blood coagulation involves multiple physiological regulatory levels, including key steps such as platelet activation, the coagulation cascade, and fibrin formation. When blood vessels are damaged, initial vasoconstriction occurs, slowing blood flow to reduce blood loss. Platelets at the injury site become activated, adhere to the damaged area, and aggregate to form an initial platelet plug—a process known as primary hemostasis, which provides only temporary hemostatic effects [14]. The secondary hemostasis is formed by the combination of fibrin polymers and platelet plugs, creating a firm and durable thrombus that occludes the wound for an extended period to achieve hemostasis.

The entire coagulation cascade consists of two main pathways: the intrinsic pathway (contact activation) and the extrinsic pathway (tissue factor). Both pathways initiate the common pathway by activating Factor X. Activated thrombin converts soluble fibrin into insoluble fibrin, which then binds to platelet plugs to form a stable thrombus, effectively halting blood flow [15].

The hemostatic principle of electrocautery in surgical procedures is primarily based on thermal effects, achieving hemostasis by heating tissues. During surgery, the electrocautery device delivers high-frequency electrical currents or laser waves, which act on the target tissue and cause a rapid localized temperature increase. Specifically, the high temperature evaporates water within tissue cells, leading to cellular desiccation and coagulation as moisture is lost. This process not only directly disrupts cellular structures but also induces contraction of surrounding vascular smooth muscle, reducing blood flow and enabling rapid hemostasis. Simultaneously, the thermal effect activates the coagulation cascade, promoting platelet aggregation and activation to form thrombi, further sealing the wound. Crucially, this thermal coagulation method offers high precision, effectively controlling bleeding without damaging adjacent healthy tissues. The use of electrocautery in surgery significantly reduces intraoperative blood loss, shortens operative time, and enhances procedural safety and efficacy. Consequently, electrocautery is widely employed in various surgical interventions, such as tumor resection and visceral organ surgeries, to ensure smooth procedures and favorable postoperative recovery outcomes.

Additionally, advancements in temperature control technology for surgical electrocautery continue to improve, minimizing excessive thermal damage during hemostasis and further enhancing safety and effectiveness. For instance, modern electrocautery devices may incorporate intelligent temperature monitoring systems that track real-time temperature fluctuations and automatically adjust power output to maintain optimal hemostatic conditions. Such high-tech applications not only improve hemostatic performance but also reduce complication risks, thereby elevating the overall success rate of surgical procedures.

3 TRADITIONAL HEMOSTATIC TECHNIQUES

Traditional hemostatic methods for combat trauma primarily include direct manual pressure, mechanical compression devices like tourniquets, and passive hemostatic dressings. Digital pressure hemostasis is a method where, after bleeding occurs, fingers are used to compress the relatively superficial portion of an artery above a bony surface (proximal end) to block blood flow and achieve rapid hemostasis. This technique is operationally simple and suitable for hemostasis in areas such as the head, neck, shoulders, and limbs, but it is difficult to maintain for extended periods [16]. Traditional hemostatic equipment such as hemostatic gauze and hemostatic cotton (subtypes of hemostatic dressings) is primarily used for packing and compression, with limited hemostatic effects and lacking other hemostatic mechanisms, thus making them difficult to meet the hemostatic demands of traumatic massive hemorrhage.

Among mechanical devices, the tourniquet, as a cornerstone technique for controlling massive hemorrhage in combat trauma, has a history of application dating back to the Hippocratic era of ancient Greece. Modern war medicine data indicate that standardized use of tourniquets can reduce mortality from extremity arterial injuries from 23.8% to 4.1%, with individual soldier compliance rates exceeding 92% [17]. The core mechanism lies in completely blocking distal arterial blood flow through mechanical compression (pressure threshold >250 mmHg), creating an “ischemia-hemostasis time window”. The Tactical Field Assessment, as a standardized procedure for hemorrhage identification, requires systematic visual and tactile examination of 12 key anatomical points such as the neck, axilla, and groin to promptly identify life-threatening bleeding signs. These include traumatic limb amputation, arterial pulsatile bleeding, blood saturation of clothing (>500 ml/m²), and compensated shock manifestations (altered mental status, systolic blood pressure <90 mmHg). This assessment system incorporates tourniquet application timeliness into the golden treatment standard: completing “High-and-Tight” binding within 1 minute of contact with the casualty achieves optimal outcomes, with a 7.3% survival rate decrease for every 10-minute delay [12].

Based on the anatomical characteristics of the injury, traditional hemostatic devices are classified into four major categories: 1) extremity tourniquets: these devices employ a dual-loop windlass design (e.g., Combat Application Tourniquet Generation 7), achieving self-locking pressure via nylon self-adhesive straps, and are suitable for controlling proximal humerus/femur hemorrhage; 2) junctional tourniquets: this category includes devices such as the SAM® Junctional Tourniquet (SJT), which utilizes rigid pressure plates to apply trans-soft-tissue compression on the iliac artery; 3) hemostatic dressings: these comprise materials like chitosan gauze (Celox®) and kaolin dressing (QuikClot®), which achieve wound hemostasis through charge adsorption or coagulation factor activation; 4) auxiliary hemostatic materials: these are substances such as hemostatic powder (WoundStat®) and fibrin glue (Tisseel®), used for supplementary hemostasis in specialized wound types [18].

Taking extremity tourniquets as an example, their technological evolution has undergone three generations of innovation: 1) the first generation (simple cloth bandage): it was widely used during the Vietnam War, with a secondary injury rate as high as 34%; 2) the second generation (elastic rubber tourniquet): it was deployed during the Gulf War, increasing hemostasis success rate to 79%; 3) the third generation (composite self-locking bandage): it was popularized after the Iraq War, utilizing polyester/nylon composite webbing and ratchet buckle, achieving a hemostasis success rate >96% and limb salvage rate of 89% [19].

However, traditional hemostatic techniques still exhibit significant limitations: 1) the time-efficacy-injury paradox: tourniquet use exceeding 2 hours can lead to nerve damage (incidence rate of 19.7%) and rhabdomyolysis; 2) anatomical blind zones: this refers to the inefficient control of subclavian artery and pelvic hemorrhage, with a success rate below 65%; 3) environmental sensitivity: dressing adhesion decreases by 37% in high-humidity environments, while low temperatures reduce thrombin activity by 52% [20].

4 NEW HEMOSTATIC MATERIAL

The rapid development of materials science has driven the innovation of hemostatic dressings, progressively replacing traditional hemostatic powders with advanced forms like hemostatic sponges and gels. Hemostatic materials can be roughly divided into three categories: polysaccharides, inorganic minerals, and proteins, each represented by various advanced formulations [21]. Polysaccharide-based materials, particularly chitosan, are widely explored. Chitosan is a natural polysaccharide extracted from shrimp and crab shells, possessing biocompatibility, biodegradability, non-toxicity, and antibacterial properties [22]. These characteristics make it an ideal absorbable hemostatic material, capable of reducing immune responses, minimizing complications, avoiding infections caused by long-term retention, and promoting wound healing. The XStat hemostatic injector is a 3 cm diameter polycarbonate syringe containing a hemostatic sponge. The XStat injectable hemostatic device achieves hemostasis through the rapid expansion mechanism of chitosan-coated cellulose sponges (**Figure 1**), providing a novel solution for addressing

Figure 1. XStat hemostatic injector. The device consists of a syringe (for delivering sponges) and chitosan-coated cellulose sponges (core hemostatic component). The sponges expand rapidly after contact with blood to form a pressure matrix.

deep non-compressible hemorrhage. This device developed by the U.S. military injects 92-100 chitosan-coated cellulose sponges (9 mm diameter × 4.5 mm height) into the wound cavity, which expands 12-fold within 12 seconds to form a three-dimensional pressure matrix. The preclinical trial of XStat was conducted using the U.S. military standard porcine femoral artery injury model (creating a 6 mm diameter hole in the porcine femoral artery), comparing its hemostatic effect with that of standard hemostatic dressings [23]. The results showed: In terms of time required to complete hemostasis, the XStat 30 group [(1.1±0.3) min] was significantly shorter than the conventional standard wound dressing group [(4.6±0.5) min]; regarding the amount of blood loss (measured from 45 seconds after creating the femoral artery hole until the end of the 6-hour observation period), the XStat group [(21.7±17.5) ml] had significantly less blood loss than the conventional standard wound dressing group [(28.3±18.1) ml] [6, 10]. In both groups, all trial pigs achieved 100% hemostasis and 100% survival. These results indicate that both XStat and standard hemostatic dressings have good hemostatic effects, but the former achieves hemostasis more rapidly.

However, the cellulose matrix of this device is non-degradable, and delayed removal will increase the incidence of granuloma formation to 17%, while high humidity environments reduce its expansion rate by 29% [19, 20].

Inorganic minerals, such as kaolin dressings, accelerate coagulation through the factor XII contact activation pathway but are limited by increased risk of failure in patients with trauma-induced coagulopathy [24]. This technical difference highlights the need to strike a balance in the development of hemostatic materials between the selectivity of coagulation mechanisms (factor-dependent/independent), material degradability, and environmental adaptability. The physical hemostatic function of kaolin is similar to that of zeolite-based materials, as both belong to aluminosilicate minerals capable of absorbing moisture from blood, concentrating clotting factors, and rapidly initiating the coagulation system [25]. Since kaolin primarily functions by activating the endogenous coagulation pathway, its hemostatic effect significantly diminishes when massive blood loss leads to substantial depletion of coagulation factors and secondary coagulation dysfunction [26]. The self-developed emergency chitosan hemostatic sponge, through positive charge-mediated erythrocyte aggregation and combined with the super absorbency (water absorption ratio ≥40 g/g) of polyacrylic resin dressings, can synergistically control deep laceration bleeding [27]. The rapid hemostatic powder "Blood Shield" uses modified zeolite as its core component. Through a dual mechanism of calcium ion release and wound charge adsorption, it achieves control of arterial and venous bleeding within 30 seconds (hemostatic efficacy rate >94%). It has been standardized as essential individual first-aid equipment [28].

Protein-based materials like fibrin glue (Tisseel®) are also used (See **Table 1**). Furthermore, composite materials combining different mechanisms are promising. The military transformation potential of traditional Chinese medicine hemostatic materials is gradually becoming evident. For instance, based on research guided by traditional medical theories, a hemostatic agent composed of *Bletilla striata* and zeolite was developed by optimizing the components. The formulation combines *Bletilla striata* lyophilized powder (the non-polysaccharide fraction of the 80% ethanol extract of *Bletilla striata*) with nano-zeolite at a 3:7 ratio, reducing coagulation time to 1.8±0.3 minutes. Additionally, coating the zeolite with *Bletilla striata* polysaccharides lowered the peak heat generation (ΔT decreased from 42 °C to 31 °C), resulting in a reduction of burn incidence from 28% to 6% [31]. Meanwhile, the 37-Chitosan synergistic system demonstrates advantages in multi-target regulation. Panax notoginseng saponins (R1+Rg1≥8%) can activate platelet GPIIb/IIIa receptors, and synergize with chitosan to promote fibrin cross-linking, thereby achieving significant effects in combat wound repair. Animal experimental results indicate that the wound healing rate is 37% faster than that of traditional dressings [30, 32].

Our military has achieved a series of breakthroughs in the field of hemostatic material development, establishing a multidimensional technical system that encompasses physical occlusion, biochemical activation, and composite mechanisms. The independently developed emergency chitosan hemostatic sponge induces red blood cell aggregation through positive charge mediation, while the superabsorbent polyacrylic resin dressing (water absorption ratio ≥40 g/g) synergistically controls deep cavity bleeding [27].

Another notable advancement is the ultrafast self-gelling/wet-state powdered adhesive developed by the research team of Bian Liming at The Chinese University of Hong Kong. It has

Table 1. Intergenerational comparison of hemostatic material technologies for combat trauma

Index	Traditional hemostatic techniques (2010-2018)	Novel hemostatic techniques (2019-2024)	Improvement extent/ innovative mechanism	Literature support
Representative equipment	Hemostatic gauze; hemostatic cotton; CAT Gen7 tourniquet; QuikClot kaolin dressing	Blood shield zeolite powder; Self-gelling adhesive; Bletilla-zeolite composite; SafeNing granules (gauze); XStat sponge syringe; Fibrin-based materials	Material compositing (biological + synthetic); Functional integration (hemostasis + repair)	[17] vs [28]
Core mechanism	Mechanical compression-induced FXII-dependent coagulation activation	Charge adsorption/self-adhesion; Non-coagulation factor-dependent mechanisms	Reduced risk of TIC failure (↓3.2-fold)	[27] vs [28]
Hemostasis time	Limb tourniquet: 45-90 s; Kaolin dressing: 2.3±0.4 min	Zeolite powder: ≤30 s; Self-gelling: ≤60 s	Timeliness improvement (58%-75%)	[17] vs [28] [29]
Secondary injury rate	Tourniquet nerve injury: 19.7%; Kaolin deep failure: 43%	Compound agent burn rate: 6%; Self-gelling agent: no thermal damage	Histocompatibility optimization (↑67%)	[18] [19] vs [30]
Environmental adaptability	High humidity failure rate (↑37%); Low-temperature coagulation activity (↓52%)	Self-gelling seawater adhesion >85%; Composite agent plateau CT <2.1 min	Extreme condition stability (↑89%)	[19] vs [29] [31]
Individual adaptability	Tourniquet volume >200 cm³; Requirement of pressure re-check	A self-gelling system (<85 cm³) with intelligent pressure maintenance (±2 mmHg)	Portability (↑58%); Operational fault tolerance (↑73%)	[19] vs [29]
Cost (USD/piece)	Tourniquets: $45-50; Kaolin dressing: $75-80	Self-gelling adhesive: $15-20; Compound agent: $35-40	Unit cost reduction (↓67%-83%)	[19] vs [29]

Note: CAT Gen7, Combat Application Tourniquet, Generation 7; FXII, Coagulation Factor XII; TIC, Trauma-Induced Coagulopathy; CT, Coagulation Time; USD, United States Dollar.

broken through the limitations of traditional formulations, exhibiting superior wet-state adhesion that achieves a bonding strength exceeding 50 kPa within one minute on hydrated wound surfaces. Simultaneously, its excellent topological adaptability enables effective filling of irregular wound cavities, with a compatibility rate exceeding 92%. Additionally, this material demonstrates significant cost advantages, with the production cost per unit dose being only 17% of commercially available hemostatic agents [29]. After undergoing tactical environment simulation tests, the material maintained stable hemostatic efficacy under seawater immersion and high-altitude low-pressure conditions.

This evolution in technical approaches highlights the critical need for balancing selectivity in coagulation mechanisms (factor-dependent vs. independent), material degradability, and environmental adaptability in the development of next-generation hemostatic materials. **Table 1** summarizes the advantages and disadvantages of different generations of hemostatic materials for combat trauma.

5 INTELLIGENT HEMOSTATIC DEVICE

Currently, the cuff-inflated pneumatic tourniquet is considered the optimal choice with high hemostatic efficiency, optimal patient comfort, and the lowest complication rates. It consists of an inflatable balloon, a pressure-release valve, and a self-adhesive inflatable cuff. Its primary drawback is skin injury. Throughout the entire application process of the tourniquet, any minor issue leading to improper use can result in skin damage, such as redness, bruising, or necrosis. Additionally, the device's susceptibility to damage, low reusability, and the difficulty for non-medical professionals to accurately control cuff pressure have prevented its widespread frontline deployment. During the research and development process, if intelligent elements could be integrated, such as enabling a "one-touch" automatic pressurization function and utilizing ultrasound or tactile sensors to automatically assess arterial compression, it would not only ensure distal blood supply while achieving hemostasis but also automatically calculate hemostasis duration and prompt timely release [33]. In this way, military medical personnel can allocate more time to rescuing other casualties.

The intelligent design has endowed this device with immense imaginative potential. In the future, inflatable hemostatic devices could be pre-assembled in junctional areas such as the neck, axillae, and groin of combat uniforms or protective gear, without interfering with soldiers' tactical movements during operations. In the event of major limb hemorrhage, a "high and tight" hemostatic effect can be achieved by triggering a switch; for junctional zone bleeding, effective compression of the bleeding area can be realized. Beyond mechanical function, research suggests integrating casualty identification systems and BeiDou positioning systems onto new-generation tourniquets [34]. The integration of intelligent elements will enable

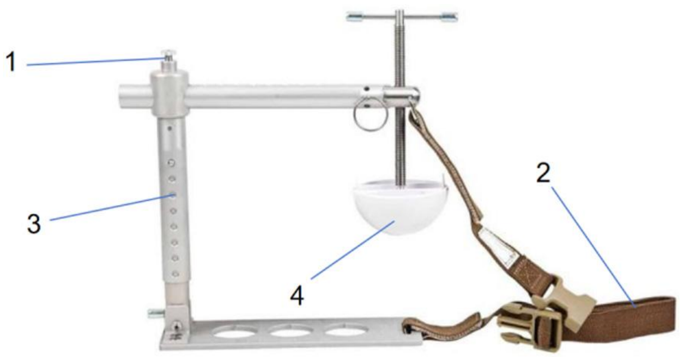


Figure 2. Combat ready clamp. 1, T-shaped threaded adjustable rod; 2, Fixing belt; 3, C-shaped steel bracket; 4, Plastic compressor.

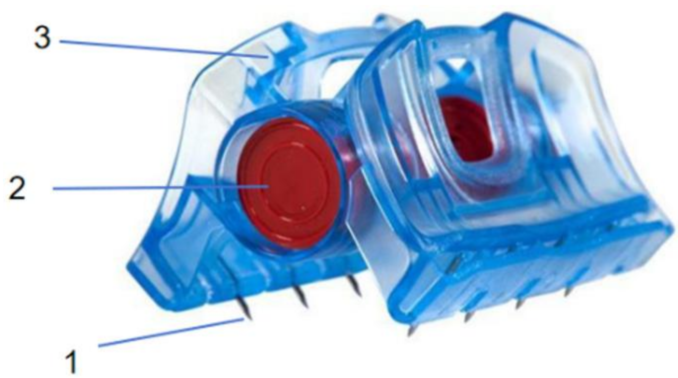


Figure 3. iTClamp hemostatic clamp. 1, Four pairs of opposing needle set; 2, Ergonomic clip body; 3, Anti-slip contact surface.

tourniquets to play an even more crucial role in modern warfare.

Traditional emergency hemostatic methods often require medical personnel to dress wounds in hazardous environments, undoubtedly increasing the risks of on-site rescue operations and even potentially endangering the safety of the rescuers themselves. Consequently, employing unmanned rescue equipment such as robotic arms to replace manual intervention holds critical significance for saving lives, alleviating the workload pressure on medical staff, and mitigating personnel shortages in healthcare. Several other widely used hemostatic devices and methods currently in practice are introduced below.

The combat-ready hemostatic clamp (**Figure 2**) weighs approximately 700 g and consists of a C-shaped steel support frame and a plastic pressure disk, serving as a clamp-type hemostatic device. This apparatus can directly apply pressure to wounds or junctional areas, compressing blood vessels to achieve effective hemostasis, making it particularly suitable for amputated limb hemorrhage control [35]. The combat ready clamp is the first junctional hemorrhage control device issued to foreign military forces, demonstrating relatively ideal hemostatic efficacy. However, it requires considerable time for assembly and is prone to detachment during transport [16].

The iTClamp hemostatic clamp (**Figure 3**) is manufactured by the Canadian company iTraumaCare. Its appearance resembles that of a common clip, weighing 37.1 grams, and is suitable for hemostasis in limbs, armpits, groin, and head/neck regions. Compared to traditional instruments, this hemostatic device employs an innovative approach by approximating wound edges to form an internal hematoma, thereby compressing damaged blood vessels and effectively controlling bleeding [36]. Unlike conventional hemostatic methods, the iTClamp design not only simplifies the operational procedure but also enhances comfort at the wound site, preventing excessive pain caused by traditional approaches for patients. In 2019, the updated U.S. Tactical Combat Casualty Care guidelines still recommended iTClamp for controlling major hemorrhage [34]. The iTClamp approximates wound edges through four pairs of opposing needles, utilizing the hydrostatic pressure generated by the hematoma to further control bleeding [16]. This hemostatic method not only does not increase pain at the wound site, but it also does not cause tissue necrosis even with prolonged use. Shang Z et al. employed a robotic arm to manage limb hemorrhage in casualties, where under the constraint of an external rigid structure, effective hemostasis was achieved through radial deformation of the airbag to compress the limb [37].

The junctional emergency treatment tool (JETT) (**Figure 4**) consists of a pelvic belt and two T-shaped handles with pressure pads, enabling bilateral compression hemostasis. By rotating the T-shaped handles to apply pressure and secure them, rapid hemostasis can be achieved in a short time (average 15 seconds). However, proficient operation is required; otherwise, it may result in inadequate hemostatic effects or equipment damage [33]. In addition, JETT can also be used for the fixation of pelvic fractures.

The SJT (**Figure 5**) is manufactured by the American company Sam Medical Products. It resembles the JETT in appearance, weighs 450 grams, and differs from the JETT in that the SJT is equipped with two inflatable cuffs. It can also be used for pelvic fractures, with a maximum usage duration of 4 hours [38]. The SJT was designed with ergonomic considerations, allowing it to fit snugly against the injured limb during use to ensure effective hemostasis. Its operation is simple, requiring medical personnel only to inflate the cuff using the accompanying air pump device to quickly achieve hemostasis. Additionally, the SJT is equipped with a pressure monitoring system that can track cuff pressure in real time, ensuring stable pressure during hemostasis and preventing tissue damage from excessive compression.

High-frequency electrocoagulation technology is a minimally invasive technique that utilizes the heat generated by high-fre-

Figure 4. Junctional emergency treatment tool. 1, Nylon waistband; 2, Two T-shaped adjustable handles with pressure pads.

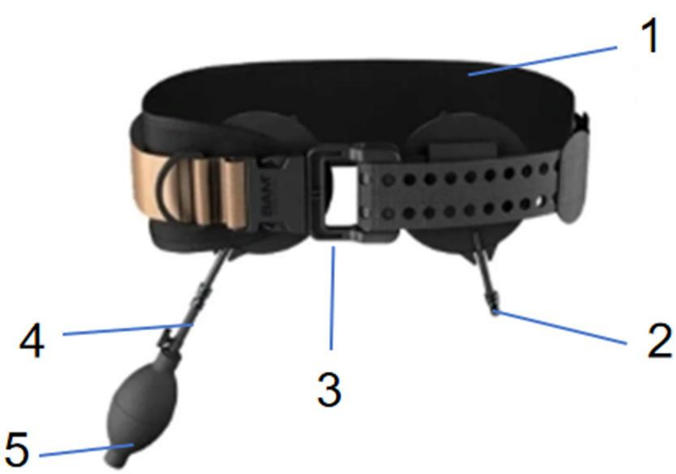


Figure 5. SAM® junction tourniquet. 1, Pressure pad expander; 2, Auxiliary Frenulum; 3, Pressure pad; 4, Pressure regulator; 5, Inflatable balloon.

quency current to coagulate tissues, thereby achieving therapeutic purposes such as cutting, hemostasis, and ablation. Due to its minimally invasive, safe, and effective characteristics, it has been increasingly widely applied in clinical practice. However, existing high-frequency electrocoagulation devices adopt a structure combining a main unit with a handpiece, which is bulky, accessory-intensive, and cumbersome to use. Common electrocoagulation handpieces also have a limited hemostatic area and low efficiency. Consequently, current high-frequency electrocoagulation hemostasis techniques cannot meet the demands of combat trauma care, such as large-area traumatic bleeding, mass casualty treatment, and time-sensitive emergency scenarios.

As a representative of our military's new hemostatic equipment, the portable field high-frequency hemostatic device achieves rapid hemostasis through the thermal coagulation effect of high-frequency current (2 MHz). Its core innovation lies in the tactical adaptation of traditional high-frequency electrosurgical technology. The device employs a bipolar coagulation mode, where precision forceps electrodes with a spacing of ≤3 mm act precisely on the bleeding site, raising local tissue temperature to 80-100 °C within 10-15 seconds to induce vascular closure. Simultaneously, the depth of thermal damage is controlled within 1.2 mm, significantly reducing the risk of tissue necrosis. The total weight of the unit is <1.5 kg, equipped with a lithium-ion battery and direct current/alternating current conversion module, enabling continuous operation for 45 minutes in environments without a mains power supply, and adaptable to complex battlefield conditions such as high-altitude, humid, and hot climates. Clinical tests demonstrate an 88% control rate for deep non-compressible hemorrhage, making it particularly suitable for irregular wounds such as shrapnel embedding injuries [39]. However, the device still requires operation by professionals, and its cost is much higher than traditional tourniquets. In the future, the integration of impedance sensing and artificial intelligence positioning modules could further enhance individual soldier adaptability.

6 CHALLENGES AND FRONTIER DIRECTIONS IN HEMOSTASIS FOR COMBAT TRAUMA

The core challenge faced by battlefield trauma hemostasis technology lies in striking a balance between efficiency, safety, and battlefield adaptability. Although traditional hemostatic methods such as tourniquets and kaolin dressings are widely used in combat settings, they struggle to overcome several critical limitations. For instance, mechanical compression may lead to nerve injuries, with an incidence rate as high as 19.7%. Meanwhile, the failure rate for deep wound hemostasis can reach 43%. Additionally, the performance degradation of these conventional methods in low-temperature or high-humidity environments cannot be overlooked. Research indicates that coagulation activity may decrease by up to 52%.

Although the novel high-frequency electrocoagulation technology can achieve rapid hemostasis through thermal effects, with the entire process taking only 8-12 seconds, it still faces challenges such as thermal damage control (thermal damage depth ≤1.5 mm), bulky equipment (weight exceeding 5 kg), and high costs (unit price approximately $1,200) [33]. Furthermore, extreme battlefield environments (such as high-altitude hypoxia or high-salt fog in maritime areas) impose even more stringent requirements on the stability of hemostatic materials. Current technologies have yet to fully meet the demands for mass casualty treatment across diverse scenarios and varying injury types.

7 FUTURE TRENDS

Under the current circumstances, innovative concepts such as intelligence, integration, lightweight design, miniaturization, recyclability, high comfort, strong stability, and low cost will provide significant reference value for the development of novel hemostatic medical devices. The future trend of hemostatic technology may focus on the application of electromagnetic induction principles—for instance, utilizing non-contact energy transfer to achieve safe and efficient new hemostatic equipment. Such devices can achieve complete isolation between the human body and the power source, fundamentally eliminating risks that traditional hemostatic methods may pose to both patients and medical staff. In contrast, traditional surgical electrocautery relies on human impedance as part of the circuit, employing high-frequency, high-voltage currents. This not only increases the risk of electric shock and burns for patients but also exposes them to direct electric shock in case of equipment failure.

Specifically, the new equipment will adopt an innovative safety design aimed at ensuring hemostatic efficiency while improving operational safety. This not only reduces surgical risks but also enhances applicability in extreme battlefield environments, better meeting the demands of modern warfare for hemostatic technology. With continuous technological advancements, future hemostatic devices will evolve toward greater intelligence, portability, and efficiency. For example, the development of nanotechnology enables micro-robots to perform precise hemostatic operations within the human body. In the future, nanorobots may be utilized to repair damaged blood vessels and control bleeding or achieve localized drug release to promote wound healing and hemostasis. Progress in bioprinting technology makes it possible to print tissues with vascular systems, which holds significant implications for hemostatic treatment and tissue repair in combat-wounded patients, providing more reliable solutions for battlefield medical care.

8 DISCUSSION AND CONCLUSIONS

The advancement of battlefield trauma hemostasis technology represents a pivotal component of modern military medicine for addressing the challenges posed by high-energy destructive weapons. From the mechanical compression of traditional tourniquets to the precise thermal effects of high-frequency electrocoagulation, and from portable intelligent hemostatic devices for individual soldiers to the integration of unmanned rescue systems, hemostatic technologies have progressively transcended single-function limitations, ushering in an era of intelligent multi-mechanism synergy and multi-scenario adaptability.

Currently, military innovation in hemostasis technology must not only resolve the balance between efficiency and safety but also overcome performance degradation in extreme environments, resource constraints in mass casualty care, and the prohibitive costs hindering large-scale deployment. This technological leap will not only save more lives on the battlefield but also reshape the medical support system, providing strategic reinforcement for sustaining combat readiness.

Ultimately, only through deepening civil-military integration and strengthening international collaboration can we surmount the challenges of cost and large-scale application, truly achieving the transition of hemostasis technology from “passive response” to “active defense”.

DECLARATIONS

Author contributions

Xinying Shi: Writing - Original Draft, Investigation; Haipo Cui: Conceptualization, Supervision, Writing - Review and Editing; Yuan Yao: Literature Review, Data Curation, Proofreading.

Funding

This work was supported by the State Key Laboratory of Systems Medicine for Cancer (KF2407-93).

Data availability

Not applicable.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

Not applicable.

REFERENCES

[1] Eastridge BJ, Mabry RL, Seguin P, Cantrell J, Tops T, Uribe P, et al. Death on the battlefield (2001-2011): Implications for the future of combat casualty care. J Trauma Acute Care Surg. 2012 Dec;73(6 Suppl 5):S431-S437. https://doi.org/10.1097/TA.0b013e3182755dcc

[2] Rich NM, Baugh JH, Hughes CW. Acute arterial injuries in vietnam: 1,000 cases. J Trauma. 1970 May;10(5):359-369. https://doi.org/10.1097/00005373-197005000-00001

[3] Guo W, He Y. Management of traumatic vascular injury in war wounds. Chin J Pract Surg. 2020;40(12):1352-1355. https://doi.org/10.19538/j.cjps.issn1005-2208.2020.12.02

[4] Sohn VY, Arthurs ZM, Herbert GS, Beekley AC, Sebesta JA. Demographics, treatment, and early outcomes in penetrating vascular combat trauma. Arch Surg. 2008 Aug;143(8):783-787. https://doi.org/10.1001/archsurg.143.8.783

[5] Kragh JF Jr, Walters TJ, Baer DG, Fox CJ, Wade CE, Salinas J, et al. Survival with emergency tourniquet use to stop bleeding in major limb trauma. Ann Surg. 2009 Jan;249(1):1-7. https://doi.org/10.1097/SLA.0b013e31818842ba

[6] Niu A, Ma H, Chen Z, Zhu X, Deng J, Luo Y. Enlightenment of the combat casualty care training in the US to Chinese military mobile medical team. J Traumat Surg. 2023;25(6):417-422. https://doi.org/10.3969/j.issn.1009-4237.2023.06.004

[7] Zong Z, Zhang L, Qin H, Chen S, Zhang L, Yang L, et al. Expert consensus on assessment and diagnosis of combat injuries of PLA. Med J Chin PLA. 2018;43(3):181-188. https://doi.org/10.11855/j.issn.0577-7402.2018.03.01

[8] Chen C, Zhang B, Feng R, Huang Z. Practical training for combat casualty care in field medical services. Med J Chin PAP. 2023 Mar;34(3):271-273. https://doi.org/10.14010/j.cnki.wjyx.2023.03.003

[9] Flecha I, Naylor JF, Schauer SG, Curtis RA, Cunningham CW. Combat lifesaver-trained, first-responder application of junctional tourniquets: A prospective, randomized, crossover trial. Mil Med Res. 2018 Sep 13;5(1):31. https://doi.org/10.1186/s40779-018-0178-1

[10] Shock and Sepsis Society of Chinese Research Hospital Association, War Wound and Trauma Society of the Chinese People's Liberation Army (PLA). Chinese expert consensus on diagnosis and treatment of traumatic multiple organ dysfunction syndrome. Chin Crit Care Med. 2022 Mar 28;34(3):225-238. https://doi.org/10.3760/cma.j.cn121430-20220118-00076

[11] Hickman DA, Pawlowski CL, Sekhon UDS, Marks J, Gupta AS. Biomaterials and advanced technologies for hemostatic management of bleeding. Adv Mater. 2018 Jan 25;30(4):10.1002/adma.201700859. https://doi.org/10.1002/adma.201700859

[12] Gurney JM, Stern CA, Kotwal RS, Cunningham CW, Burelison DR, Gross KR, et al. Tactical combat casualty care training, knowledge, and utilization in the us army. Mil Med. 2020 Feb 19;185(Suppl 1):500-507. https://doi.org/10.1093/milmed/usz303

[13] Wei M, Qin X. Characteristics and enlightenment of tactical combat casualty care in the U.S. military. People's Mil Surg. 2020;63(12):1183-1185. https://doi.org/10.3969/j.issn.1000-9736.2020.12.013

[14] Tan C. Study on preparation and properties of CS/PVA-alginate composite hemostatic dressing [master's thesis]. Shanghai: Donghua University; 2020. 81 p.

[15] Cornelius B, Campbell R, McGauly P. Tourniquets in trauma care: A review of application. J Trauma Nurs. 2017 May-Jun;24(3):203-207. https://doi.org/10.1097/jtn.0000000000000290

[16] Xing Y, Li H. Research progress on hemostasis in naval battle injuries. J Navy Med. 2024 Apr;45(4):351-354. https://doi.org/10.3969/j.issn.1009-0754.2024.04.007

[17] Kotwal RS, Montgomery HR, Miles EA, Conklin CC, Hall MT, McChrystal SA. Leadership and a casualty response system for eliminating preventable death. J Trauma Acute Care Surg. 2017 Jun;82(6S Suppl 1):S9-S15. https://doi.org/10.1097/ta.0000000000001428

[18] Drew B, Montgomery HR, Butler FK Jr. Tactical Combat Casualty Care (TCCC) Guidelines for Medical Personnel: 05 November 2020. J Spec Oper Med. 2020 Winter;20(4):144-151. https://doi.org/10.55460/RBRA-WMWV

[19] Sims K, Montgomery HR, Dituro P, Kheirabadi BS, Butler FK. Management of external hemorrhage in tactical combat casualty care: The adjunctive use of xstat™ compressed hemostatic sponges: Tccc guidelines change 15-03. J Spec Oper Med. 2016 Spring;16(1):19-28. https://doi.org/10.55460/6cem-36iy

[20] Kragh JF Jr, Dubick MA, Aden JK, McKeague AL, Rasmussen TE, Baer DG, et al. U.S. military use of tourniquets from 2001 to 2010. Prehosp Emerg Care. 2015 Apr-Jun;19(2):184-190. https://doi.org/10.3109/10903127.2014.964892

[21] Sun Y, Dong W, Jing Z. Frontiers of vascular injury in war: Advanced treatment techniques. J Trauma Surg. 2024;26(3):161-166. https://doi.org/10.3969/j.issn.1009-4237.2024.03.001

[22] Alemu D, Getachew E, Mondal AK. Study on the physicochemical properties of chitosan and their applications in the biomedical sector. Int J Polym Sci. 2023 Jul 5;2023(1):5025341. https://doi.org/10.1155/2023/5025341

[23] Kragh JF Jr, Aden JK, Steinbaugh J, Bullard M, Dubick MA. Gauze vs xstat in wound packing for hemorrhage control. Am J Emerg Med. 2015 Jul;33(7):974-976. https://doi.org/10.1016/j.ajem.2015.03.048

[24] Basadonna G. Quikclot combat gauze for hemorrhage control. Prehosp Disaster Med. 2012 Apr;27(2):217. https://doi.org/10.1017/s1049023x12000490

[25] Bennett BL, Littlejohn L. Review of new topical hemostatic dressings for combat casualty care. Mil Med. 2014 May 1;179(5):497-514. https://doi.org/10.7205/milmed-d-13-00199

[26] Peng HT, Grodecki R, Rizoli S, Shek PN. A comparative study of tissue factor and kaolin on blood coagulation assays using rotational thromboelastometry and thromboelastography. Blood Coagul Fibrinolysis. 2016 Jan;27(1):31-41. https://doi.org/10.1097/mbc.0000000000000381

[27] Shi Y, Zhu H, Zhang Q, Wang Z, Du B, Zhu K, et al. Development and application research of trauma emergency hemostatic dressing. Mil Med J Southeast China. 2019 Mar;21(2):219-221. https://doi.org/10.3969/j.issn.1672-271X.2019.02.027

[28] Zhang P, Xiao N, Zhang Z, Li S. Current status and prospect of haemostatic dressing for war injury. J Trauma Surg. 2009;11(4):378-380. https://doi.org/10.3969/j.issn.1009-4237.2009.04.033

[29] Peng X, Xia X, Xu X, Yang X, Yang B, Zhao P, et al. Ultrafast self-gelling powder mediates robust wet adhesion to promote healing of gastrointestinal perforations. Sci Adv. 2021 Jun 2;7(23):eabe8739. https://doi.org/10.1126/sciadv.abe8739

[30] Dong Y, Liu X, Dong L, Wang X, Huang Y, Wang Y. Study on hemostatic effect and mechanism of polysaccharides from Bletilla striata in blood heat and hemorrhage model rats. China Pharm. 2016;27(31):4347-4350. https://doi.org/10.6039/j.issn.1001-0408.2016.31.08

[31] Da H, Jiao F, Chen S, Xia Z. The clinical application of the novel elastic pressure dressing for hemorrhagic injury in out-patient

and emergency department. Mil Med J Southeast China. 2015 May;17(3):266-268. https://doi.org/10.3969/j.issn.1672-271X.2015.03.013

[32] Zhao F, Yang X, Xu D, Dong L, Li J, Wang Y, et al. Hemostatic effect and mechanism of a non-polysaccharide fraction of Bletilla striata. Chin Pharmacol Bull. 2016 Aug;32(8):1121-1126. https://doi.org/10.3969/j.issn.1001-1978.2016.08.018

[33] Johnson JE, Sims K, Hamilton DJ, Kragh JF Jr. Safety and effectiveness evidence of SAM® junctional tourniquet to control inguinal hemorrhage in a perfused cadaver model. J Spec Oper Med. 2014 Summer;14(2):21-25. https://doi.org/10.55460/d5ps-nyqa

[34] Liu W, Li L. Foreign military medical series (129): The new US military tactical combat casualty care guidelines and related technical advances. People's Mil Surg. 2019;62(8):719-722.

[35] Guo D, Zhang R, Li T, Pan F, He W. A review of theories, techniques and equipment for first-aid hemostasis at the modern battlefield. Chin J Disaster Med. 2021 Jun;9(6):1066-1070. https://doi.org/10.13919/j.issn.2095-6274.2021.06.011

[36] Mottet K, Filips D, Logsetty S, Atkinson I. Evaluation of the it-clamp 50 in a human cadaver model of severe compressible bleeding. J Trauma Acute Care Surg. 2014 Mar;76(3):791-797. https://doi.org/10.1097/ta.0000000000000043

[37] Shang Z, Wang T, Shao S, Zhang K. Design and analysis of a rigid-soft hybrid hemostatic manipulator. Robot. 2024 Nov;46(6):683-691. https://doi.org/10.13973/j.cnki.robot.230346

[38] Jiang J, Zhang J, Li X, Luo Y, Guo J, You H. Current situation and thinking of first aid hemostatic equipment for war wound between Chinese and American militaries. J Trauma Surg. 2021;23(2):148-151. https://doi.org/10.3969/j.issn.1009-4237.2021.02.018

[39] Zhu X, Shi Z, Tang L, Zhang Z, Qi S, Wang Y, et al. Design of field portable high-frequency hemostasis device. China Med Devices. 2011;26(9):40-44. https://doi.org/10.3969/j.issn.1674-1633.2011.09.010

Progress in
Medical Devices

Zentime

REVIEW ARTICLE

AI-assisted diagnosis of myocardial hypertrophy based on cardiac MRI: A systemic review

Shimin Zhou[1], Xudong Guo[1,2], Yunli Shen[2], Qinfen Jiang[2], Xin Gong[2], Jie Ding[2], Yihong Yang[3], Guojie Xu[1], Jican Wen[1], Jingyang Niu[1]

[1]School of Health Science and Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China.
[2]State Key Laboratory of Cardiovascular Diseases and Medical Innovation Center, Shanghai East Hospital, School of Medicine, Tongji University, Shanghai 200093, China.
[3]Department of Nuclear Medicine, Shanghai East Hospital, School of Medicine, Tongji University, Shanghai 200120, China.

Corresponding author: Xudong Guo.

Address correspondence to: Xudong Guo, School of Health Science and Engineering, University of Shanghai for Science and Technology, No. 516 Jungong Road, Yangpu District, Shanghai 200093, China.
E-mail: guoxd@usst.edu.cn.

Received October 25, 2025; Accepted February 12, 2026; Published March 31, 2026

DOI: 10.61189/569607adnpiw

Abstract

Cardiac hypertrophy represents a complex pathological condition characterized by ventricular wall thickening, with diverse etiologies and substantial challenges in clinical differential diagnosis. In recent years, rapid advances in artificial intelligence (AI) techniques for CMR image analysis have provided novel technical approaches for the precise diagnosis of cardiac hypertrophy. This paper systematically reviews the research progress of CMR-based AI technologies in the diagnosis of cardiac hypertrophy, including AI diagnostic methods based on Cine-MRI sequences, T1/T2 Mapping sequences, late gadolinium enhancement (LGE) sequences, and multi-sequence fusion strategies. The review further explores the technological evolution from traditional machine learning to deep learning and their applications in differentiating normal from hypertrophic hearts, as well as in the fine classification of cardiac hypertrophy with different etiologies. Furthermore, this paper elucidates the application value of natural language processing (NLP)-based MRI report automatic parsing technology in large-scale case screening and discusses the existing challenges and potential future directions of AI in this field.

Keywords: Cardiac hypertrophy, Cardiac magnetic resonance, Artificial intelligence, Deep learning, Multi-sequence fusion

Highlights

- This review systematically summarizes the research progress of artificial intelligence technologies in the diagnosis of cardiac hypertrophy based on cardiac MRI, with a focus on AI diagnostic methods utilizing Cine-MRI, T1/T2 Mapping, late gadolinium enhancement (LGE), and multi-sequence fusion strategies.
- This review highlights the application potential and current limitations of natural language processing-based automated MRI report parsing technology for large-scale case screening and phenotypic stratification.
- This review analyzes existing challenges in AI diagnosis, including data quality, annotation consistency, and model generalization, and discusses future directions such as multicenter collaboration, multimodal data fusion, and clinical translation.

1 INTRODUCTION

Cardiac hypertrophy is a pathological condition characterized by abnormal thickening of the ventricular wall, typically manifested as a significant increase in left ventricular (LV) wall thickness. The etiologies of cardiac hypertrophy are diverse and may arise from both physiological and pathological mechanisms, including athletic cardiac remodeling, hypertensive heart disease (HHD), aortic stenosis (AS), hypertrophic cardiomyopathy (HCM), infiltrative cardiomyopathies, and various storage or metabolic disorders such as cardiac amyloidosis (CA) and Anderson–Fabry disease (AFD) [1]. Cardiac hypertrophy is associated with increased risk of heart failure, arrhythmia, cardiovascular mortality, and sudden cardiac death. Early recognition and accurate determination of the underlying etiology can substantially improve clinical outcomes, particularly in patients under 60 years of age [2]. However, the wide spectrum of causes, clinical manifestations, and disease progression make accurate diagnosis and risk stratification a persistent clinical challenge.

Imaging plays a central role in the diagnostic workflow of cardiac hypertrophy. Conventional echocardiography is broadly accessible but is operator-dependent and provides limited quantification of myocardial tissue characteristics. Cardiac computed tomography (CT) provides excellent anatomical visualization but involves ionizing radiation and lacks dynamic functional assessment. In contrast, cardiac magnetic resonance (CMR) provides superior soft tissue resolution, precise and highly reproducible measurements (including myocardial wall thickness, chamber size, hypertrophy patterns, and systolic function), and multiparametric imaging capabilities, such as T1 mapping anPlr3 ladolinium enhancement (LGE). These advantages have established CMR as the “gold standard” for assessing the etiology of cardiac hypertrophy, the degree of fibrosis, and functional impairment, making it the most valuable imaging modality for evaluating LV hypertrophy [3].

With the growing demand for cardiac imaging and the continuous advancement of technology, CMR analysis faces several challenges associated with image acquisition, data quality, annotation, and concerns regarding patient privacy, data sharing, and institutional standardization. Moreover, conventional analysis methods often struggle to accommodate patients with complex or atypical pathological conditions [4]. Against this background, the rapid development of AI-based medical imaging analysis presents a promising solution that provides new approaches for improving the accuracy, standardization, and efficiency of cardiac MRI interpretation [5].

Although significant advances has been made in AI-assisted diagnosis of cardiac hypertrophy, existing studies are fragmented and rarely provide a comprehensive overview focused on CMR, the “gold standard” imaging modality. Therefore, this review systematically examines the research progress of CMR-based AI technologies in diagnosing cardiac hypertrophy and explores the current challenges and future directions of AI in intelligent diagnostic support for cardiac hypertrophy disease.

2 RESEARCH PROGRESS IN AI-ASSISTED DIAGNOSIS OF CARDIAC HYPERTROPHY BASED ON CMR MODALITIES

2.1 AI Diagnosis of cardiac hypertrophy based on cine-MRI

Cine-CMR acquires continuous dynamic images throughout the cardiac cycle using the steady-state free precession (SSFP) sequence, which enables detailed observation of morphological changes of heart with high temporal and spatial resolution. Cine-CMR has become the gold standard for cardiac structural and functional assessment. For the diagnosis of cardiac hypertrophy diseases, Cine-MRI can precisely identify the location, extent, and pattern of ventricular wall thickening, distinguish physiological from pathological hypertrophy, differentiate myocardial lesions caused by different etiologies, and assess the hemodynamic impact of the disease through cardiac function parameters, providing critical information for clinical decision-making.

2.1.1 Detection and identification of cardiac hypertrophy

In deep learning applications, Budai et al. developed an ensemble model based on 3D ResNet for automated detection of left ventricular hypertrophy (LVH) in CMR [6]. The study involved 428 LVH patients and 234 healthy controls. The ensemble model simultaneously processed short-axis and long-axis cardiac images, yielding highly precise binary classification between LVH and normal hearts, with a recall rate of 97%, demonstrating exceptional ability in identifying suspected cases. A key innovation was the introduction of image normalization through long-axis superimposition to mitigate directional noise, significantly enhancing model performance.

To address the limitation of traditional single-frame or single-branch deep learning models in adequately capturing cardiac dynamic features, You et al. proposed an innovative solution based on a dual-branch neural network integrating static morphology and dynamic motion information [7]. The architecture combined static and dynamic features and innovatively introduced a multi-task learning mechanism, where the primary task performed cardiomyopathy classification and the auxiliary task quantified cardiac motion amplitude by comparing standard deviations of images at identical positions in static and dynamic branches. On a dataset containing 78 training sequences and 26 test sequences, this method achieved 96.79% accuracy and 95.24% sensitivity for binary classification of hypertrophic obstructive cardiomyopathy (HOCM) versus normal hearts, representing improvements of 10.11% and 12.02% over ResNet50 and DenseNet, respectively. More importantly, the loss curve basically converged after 100 epochs, and the training efficiency was approximately twice that of the traditional single-branch network.

Ma et al. further explored the potential application of radiomics for early HHD diagnosis, enrolling 132 patients [8]. A total of 1,521 radiomics features were extracted from the end-diastolic (ED) and end-systolic (ES) phases of cine sequences, and three-class classification models were built using machine learning (ML) algorithms such as random forest, support vector machine (SVM), and naive Bayes. The SVM model based on the ED-ES feature subset achieved the highest accuracy (83.3%), with a macro-AUC of 0.941, and the AUCs for HHD, hypertension with normal cardiac structure and function (HWN), and healthy subjects were 0.967, 0.876, and 0.963, respectively.

In addition to the morphological classification models, automatic analyses of cardiac strain parameters have shown unique diagnostic value. Abràmoff et al. used commercial AI software (SuiteHEART) for fully automated LV global longitudinal strain (GLS) analysis in 111 CMR examinations (70 AS, 41 healthy controls) [9]. The results showed that the AI-driven CMR GLS measurements were strongly associated with echocardiographic GLS and exhibited smaller standard deviations, indicating higher accuracy and reproducibility.

2.1.2 Differentiation of cardiac hypertrophy phenotype from other cardiomyopathies

Studies in this domain employ multi-class classification frameworks to position cardiac hypertrophy within the broader spectrum of cardiomyopathies, utilizing both traditional radiomics-based machine learning (ML) and deep learning approaches to achieve phenotypic differentiation. Zhang et al. analyzed 283 patients from two public databases (ACDC and M&Ms) to establish a multi-classification model based on radiomics features [10]. The research team extracted 21 feature subsets from ED-ES images of the LV, right ventricle, and myocardium, testing nine ML algorithms, including random forest, SVM, and logistic regression. Among 90 model combinations, the random forest model using the minimum redundancy maximum relevance (mRMR) feature selection achieved optimal performance, with a three-class classification accuracy of 91.2% and a macro-average AUC of 0.947. Class-specific AUCs were 0.938, 0.966, and 0.936 for healthy controls, dilated cardiomyopathy (DCM), and HCM, respectively. Five-fold cross-validation demonstrated an average accuracy of 83.0%, sensitivity of 80.4%, and specificity of 90.6%.

The development of deep learning technologies has enabled end-to-end analysis of Cine-MRI images, automatically learning hierarchical representations and mapping raw images to diagnostic outcomes. Germain et al. retrospectively analyzed 534 patients (209 normal, 175 HCM, 150 DCM) to evaluate the performance of several pre-trained convolutional neural networks (CNNs) for cardiomyopathy classification [11]. Transfer learning strategies employed VGG16, ResNet50V2, InceptionResNetV2, and DenseNet201 pre-trained on ImageNet as backbone networks, with final layers fine-tuned to adapt to cardiac imaging. The result shows that the dual-input VGG model achieved 98.2% accuracy in six-fold cross-validation, substantially higher than the accuracy of single-frame input model. Grad-CAM heatmap analysis revealed that over half of misclassifications were due to the network focusing on extra-cardiac structures rather than the LV region, indicating that cardiac region segmentation preprocessing could improve model robustness.

Guo et al. systematically evaluated the performance of commercial AI software in quantitative LV function analysis in 388 patients (123 HCM, 130 DCM, 135 healthy controls) [12]. AI-based automatic analysis performed best in HCM patients, with correlations exceeding 0.901 between all four LV function parameters and manual measurements, while performance significantly declined in DCM patients, with the correlation coefficients for ejection fraction (EF) and stroke volume (SV) dropping to 0.776 and 0.645, respectively. Receiver operating characteristic (ROC) curve analysis indicated that, at optimal cut-off values, AI-derived EF parameters identified DCM patients with 92.31% sensitivity and 82.96% specificity, whereas HCM identification showed 78.05% sensitivity and 54.07% specificity, suggesting that LV volume parameters are more informative in DCM diagnosis, whereas EF is more suitable for HCM differentiation.

In small sample learning scenarios, few-shot learning (FSL) provides an innovative solution for alleviating the data scarcity problem in medical imaging. Wibowo et al. proposed a two-stage approach combining segmentation and few-shot classification using the MICCAI 2017 ACDC dataset [13]. A “2D Thickness Algorithm” was designed to convert multi-layer short-axis slices from U-Net segmentation outputs into compact 2D representations, directly extracting morphological information from segmentation results. The EfficientNetB5-UNet segmentation model achieved LV Dice Score of 96.24% ED and 89.92% ES, and the integrated few-shot classifier reached 92% accuracy on the test set, which is comparable to the traditional random forest models based on derived features. This study demonstrated the feasibility of constructing disease-specific representations from segmentation outputs, providing a lightweight solution for fast screening under data-limited conditions, with potential expansion to long-axis data and other cardiac disease subtypes.

2.1.3 Etiological differentiation of cardiac hypertrophy

Following the identification of the hypertrophic phenotype, clinical priority shifts towards precise etiological characterization, which is critical given the distinct therapeutic strategies required for different pathologies. However, differentiating between etiologies with substantial phenotypic overlap—such as HCM, HHD, and cardiac amyloidosis (CA)—remains a diagnostic challenge. Conventional visual assessment and standard quantitative metrics often lack the sensitivity to detect subtle morphological differences. Consequently, this section focuses on the application of AI in the fine-grained differential diagnosis of cardiac hypertrophy with similar macroscopic presentations.

Jiang et al. systematically applied texture analysis combined with machine learning in a single-center study of 251 patients [14]. They extracted 275 radiomics features from four-chamber view images rather than traditional short-axis frames, achieving objective differentiation between CA and HCM. Among various machine learning models, SVM achieved the highest validation accuracy of 85.2%. Gray-level non-uniformity was identified as the most discriminative single feature, reflecting greater microstructural heterogeneity in HCM myocardium due to myocyte disarray and fibrosis.

Liu et al. expanded this approach, not only extracting features from the myocardium but also, for the first time, systematically introducing 3D radiomics features of papillary muscles, to improve left ventricular hypertrophy (LVH) detection and HCM versus HHD differentiation [15]. The study enrolled 345 subjects, and 2,632 original features were extracted, selecting the six most discriminative features. For LVH detection, myocardial (MYO) features alone achieved an AUC of 0.966 with 90.4% accuracy. For HCM versus HHD differentiation, the MYO combined with papillary muscle features (MYO+PM) achieved an AUC of 0.935 (87.0% accuracy, 85% specificity), significantly superior to MYO alone (AUC=0.875) and models based solely on CMR parameters. This study demonstrated that papillary muscle morphology differs between etiologies, with concentric hypertrophy in HCM and pressure-driven hypertrophy in HHD, and that these differences are captured by shape-based radiomics features.

Chuah et al. first introduced the use of a personalized 3D+temporal LV model together with traditional ML for phenotypic analysis between HHD and HCM [16]. Based on 44 subjects, dynamic LV models were created, extracting multiparametric features including wall thickness, strain, and systolic synchrony. Statistical classification identified the most discriminating features, such as maximum end-diastolic wall thickness, longitudinal strain, and mass-to-volume ratio. SVM achieved 77% classification accuracy, improving to 94% after excluding subgroups with phenotypic overlap.

Similarly, Eckstein et al. utilized myocardial strain parameters from both atria and the right ventricle for differentiating CA, HCM, and healthy controls [17]. An SVM model achieved 90.9% accuracy and an AUC of 0.996, underscoring the diagnostic value of myocardial biomechanical features in etiological differentiation.

In deep learning applications, Diao et al. proposed a fully automated framework for segmentation and multietiology classification in 302 LVH patients [18]. Three classification models were tested: (1) Model 1 with 2D patches, (2) Model 2 with ROI-based input, and (3) Model 3 with binary myocardium masks. Model 3 performed the best and most stable, with a three-class accuracy of 77.4% in the external validation. Notably, simple binary masks performed better than raw images for generalization, suggesting that structural representations of inputs may enhance model robustness under limited data conditions.

Chen et al. focused on distinguishing HCM from Fabry cardiomyopathy, an important clinical challenge [19]. Using 215 datasets, they trained a 3D ResNet18 model on short-axis Cine images. Internal testing yielded an AUC of 91.4%, with external validation across different hospitals achieving an AUC of 0.918, indicating strong generalizability.

In a cohort of 119 CA patients and 122 LVH patients of other causes, Germain et al. employed a dual-channel VGG16 model to distinguish CA from other LVH etiologies [20]. The model analyzed diastolic and systolic cine images, significantly outperforming three experienced imaging physicians at both frame-level (AUC: 0.824 vs. 0.630) and patient-level (AUC: 0.895 vs. 0.727) classification.

Beyond differentiating common morphological diseases, deep learning based on Cine-MRI can further reveal genotype-related information. Zhou et al. pioneered the use of four-chamber view cine sequences to predict genetic mutation status in HCM patients [21]. They combined InceptionResNetV2 for image feature extraction with long short-term memory (LSTM) networks for temporal dynamics analysis, achieving an AUC of 0.80, superior to traditional clinical scoring systems. This work suggests that subtle myocardial structural and motion patterns embedded in Cine-MRI may correlate with underlying genetic defects, providing a new perspective for non-invasive genetic risk stratification. Summaries of representative studies are presented in **Table 1**.

2.2 AI diagnosis of cardiac hypertrophy based on T1/T2 mapping sequences

T1 and T2 mapping are standardized CMR quantitative tissue characterization sequences that reflect microscopic changes in the myocardial extracellular matrix and water content. Compared with conventional morphological indicators, these sequences provide earlier and more sensitive pathological information [22]. T1/T2 mapping can detect subclinical lesions, such as diffuse fibrosis and edema in a contrast-free or low-dose contrast environments, making them excellent candidates for AI diagnosis: quantitative parameters are objective and reproducible, while texture analysis of parameters can further explore tissue heterogeneity at microscopic level. In recent years, the application of deep learning and radiomics to T1/T2 mapping have demonstrated improved accuracy in the differential diagnosis of cardiac hypertrophy of varying etiologies [23].

Table 1. AI-assisted diagnosis of cardiac hypertrophy based on cine-MRI

Ref.	AI method	Sample size	Sensitivity (%)	Specificity (%)	Accuracy (%)	AUC
Zhang et al. [10]	ML	HCM: 48, DCM: 52, NOR: 123	80.40	90.60	91.20	0.936
Eckstein et al. [17]	ML	CA: 43, HCM: 20, NOR: 44	85.67	92.33	90.90	0.996
Ma et al. [8]	ML	HHD: 42, HWN: 46, NOR: 44	83.30	98.90	97.20	0.967
Germain et al. [11]	Dual VGGNet+MLP	NOR: 395, HCM: 411, DCM: 394	-	-	98.20	-
Guo et al. [12]	CNN (U-Net)	HCM: 123, DCM: 130, NOR: 135	92.31	82.96	88.60	0.932
You et al. [7]	CNN+two LSTM-based branches	HOCM: 48, NOR: 30	95.24	-	96.79	-
Jiang et al. [14]	ML	CA: 85, HCM: 82, NOR: 84	85.00	85.00	85.00	0.890
Liu et al. [15]	ML	HCM: 158, HHD: 72, NOR: 115	77.00	83.00	83.00	0.870
Budai et al. [6]	3D ResNet	HCM: 346, CA: 45, AFD: 11, EMF: 16, AS: 10, NOR: 234	96.00	-	91.00	-
Germain et al. [20]	VGGNet	CA: 119, other LVH: 122	85.70	77.60	82.50	0.895
Diao et al. [18]	CNN+RNN+SVM	CA: 53, HCM: 82, HHD: 56	-	-	77.40	-
Zhou et al. [21]	InceptionResnetV2	HCM gene positive: 98, gene negative: 100	85.71	69.57	84.31	0.840
Chen et al. [19]	3D ResNet18	AFD: 176, HCM: 70	-	-	90.90	0.910

Note: AFD, Anderson–Fabry disease; AS, Aortic Stenosis; AUC, area under the curve; CA, cardiac amyloidosis; CNN, convolutional neural networks; DL, deep learning; DCM, dilated cardiomyopathy; EMF, endomyocardial fibrosis; HHD, hypertensive heart disease; HCM, hypertrophic cardiomyopathy; HOCM, hypertrophic obstructive cardiomyopathy; LVH, left-ventricular hypertrophy; ML, machine learning; NOR, normal person.

To address multi-etiology differentiation in cardiac hypertrophy, Antonopoulos et al. proposed a machine learning classification framework based on T1 mapping radiomics [24]. 850 radiomics features were extracted from 149 subjects. Following feature stability assessment and removal of highly correlated features, a total of 84 features were selected for multinomial logistic regression modeling using a random forest algorithm for feature importance ranking. The model achieved a multiclass AUC of 0.753 in distinguishing normal, LVH, HCM, and CA, with performance notably superior to the traditional native T1 mean values across all categories. Similarly, Shi et al. performed a systematic comparison, simultaneously assessing the discriminative power of T1 and extracellular volume (ECV) mapping texture analysis for differentiating HCM, HHD, and healthy controls, compared to functional strain parameters [25]. In a cohort of 145 patients, 271 texture features were extracted from T1 and ECV maps. ECV texture analysis achieved better diagnostic performance (AUC 0.894) compared to T1 texture for distinguishing between HCM and HHD, while performance was similar for patients versus healthy controls. These findings indicate that standardized ECV characteristics provide advantages in etiological differentiation.

Traditional T2-weighted imaging is often used to detect myocardial edema, but interpretation is subjective. Huang et al. applied texture analysis to T2-weighted CMR for differentiating CA from HCM [26]. In 317 patients, a total of 837 texture features were extracted from short-axis T2 STIR sequence, with 7 optimal features selected, achieving an AUC of 0.842; accuracy dropped to 79.2% on an independent test set of 90 patients. Notably, the texture model did not differ significantly from quantitative LGE in its performance for diagnosis, suggesting that contrast-free T2 texture analysis could serve an alternative to LGE, especially in CA patients with compromised renal function.

Beyond diagnostic value, AI-derived parameters from mapping sequences have demonstrated significant prognostic potential. Hwang et al. used Myomics-Q deep learning algorithm for fully automated native T1 and ECV measurements, where they found that ECV effectively differentiated CA from other LVH etiology (AUC=0.946) and provided prognostic prediction for AL-CA patients [27]. The study identified ECV≥40% as an independent risk factor for the composite endpoint of cardiovascular death or heart failure hospitalization in AL-CA patients. Integrating automated ECV into the updated Mayo staging system greatly enhanced patient risk classification, with Integrated Discrimination Improvement (IDI) increasing by 27.9% and Net Reclassification Index (NRI) increasing by 13.8%.

2.3 AI diagnosis of cardiac hypertrophy based on LGE sequences

LGE imaging is the standard CMR technique for myocardial tissue characterization. By visualizing the retention patterns of gadolinium-based contrast agents in the myocardial interstitium, LGE can intuitively reflect the distribution of myocardial pathological remodeling, making it an important tool for distinguishing etiologies of cardiac hypertrophy [28]. However, traditional visual assessment is limited by the experience of observers, variability in the scanning parameters, and it is difficult to detect lesions in the early stage. In recent years, AI technologies have enabled automatic identification of pathological patterns from LGE images, providing more objective judgments and improving early detection rates.

Martini et al. developed a CNN model for the end-to-end analysis of LGE images in 206 patients with suspected CA, using raw short-axis, two-chamber, and four-chamber LGE views [29]. The model achieved a high AUC of 0.982 in the test set. This study showed that deep learning models could automatically capture complex image patterns of CA, including myocardial enhancement, early darkening of the blood pool, and even extracardiac features such as pericardial fluid, with results comparable to those of expert radiologists, but with higher efficiency. Notably, the deep learning model did not significantly outperform CMR features-based machine learning models, indicating that in the case of limited data, the superiority of end-to-end deep learning over feature engineering may have been exaggerated.

To address the limited information extracted from single-sequence LGE, Zhou et al. employed high-throughput radiomics to analyze LGE images of 200 biopsy-confirmed AL-type amyloidosis patients [30]. A total of 1,906 features representing texture, shape, and signal intensity distribution were extracted. An XGBoost-based diagnostic model demonstrated excellent performance in both internal and external multicenter validation. The radiomics model of the LV basal segments achieved the highest external validation AUC of 0.92, significantly superior to traditional visual assessment (AUC=0.75). More importantly, the radiomics score moderately correlated with CA disease severity, suggesting that LGE-based radiomics features can be used not only for diagnosis but also for assessing disease burden and prognosis.

AI-based LGE diagnostic methods have demonstrated good diagnostic performance in specific myocardial diseases, especially amyloidosis. However, comparisons of AI paradigms with similar diagnostic performance but different approaches indicate that, on small-scale datasets, the absolute advantage of deep learning over carefully designed feature-based models may be limited. To advance AI applications of LGE imaging in the differential diagnosis of cardiac hypertrophy, larger-scale multicenter annotated databases, multi-sequence data fusion strategies, and rigorous external validation are required, alongside evaluation of clinical implementability.

2.4 AI diagnosis of cardiac hypertrophy based on multi-sequence CMR fusion

Single-sequence image analysis often fails to fully characterize the essential features of cardiac hypertrophy, as different CMR sequences provide complementary pathological information. Cine-MRI captures dynamic cardiac function and morphological changes, LGE sequences display spatial distribution of fibrosis, and T1/T2 mapping provides quantitative tissue characterization. Multi-sequence CMR data fusion integrates complementary information from different sequences (e.g., cine imaging, LGE, and strain analysis) to construct more discriminative feature sets, thereby achieving more precise diagnostic classification and risk assessment of myocardial diseases.

Zhang et al. extracted 1,409 radiomics features from Cine and LGE sequences in 621 patients [31]. After rigorous feature screening and multivariate logistic regression optimization, 40 optimal features were selected to establish a combined model, achieving AUCs of 0.979 and 0.981 in the training and validation cohorts, respectively, with sensitivities of 95.2% and 86.3%. This demonstrates that fusion of deep image information beyond human visual discrimination can greatly improve model diagnostic accuracy, highlighting the potential of high-precision diagnostic assistance systems.

Kong et al. employed principal component analysis (PCA) to project highly correlated parameters such as left ventricular ejection fraction (LVEF), left ventricular end-systolic volume index (LVESVI), left ventricular end-diastolic volume index (LVEDVI), maximum left ventricular wall thickness (MLVWT), and global circumferential strain into an orthogonal principal component space, developing an Integrated Algorithm (IntA) [32]. In Phase I (148 patients: 75 HCM, 33 HHD, 40 controls), IntA achieved an AUC of 0.900 (83% sensitivity, 91% specificity) in LGE-positive patients and an AUC of 0.947 (100% sensitivity, 82% specificity) in LGE-negative patients. Phase II external validation of 71 patients maintained good performance (LGE-positive AUC of 0.857, LGE-negative AUC of 0.846), demonstrating robust generalization. PCA effectively addressed multiparameter collinearity through dimensionality reduction without information loss, preserving over 90% of original variance.

With the widespread application of AI in medical image processing, enhanced automated feature extraction and fusion capabilities have enabled high-resolution myocardial phenotype characterization. Lu et al. utilized deep convolutional neural networks (DCNN) for automated CMR segmentation, generating 216 fine-segment LV thickness features [33]. Logistic regression, SVM, and multilayer neural networks were used for adverse event prediction in cardiac sarcoidosis patients. Features obtained from automated processing improved machine learning model accuracy compared with manually delineated features: logistic regression from 0.75 to 0.78, SVM from 0.74 to 0.78, and multilayer perceptron (MLP) from 0.76 to 0.77.

Three-dimensional myocardial deformation analysis (3D-MDA) achieved more refined multi-sequence fusion through cross-modal standardization. Satriano et al. reconstructed 3D dynamic meshes from Cine-MRI multi-planar images using optical flow registration, calculating 917 structural and deformation parameters per American Heart Association (AHA) segment, including peak strain amplitude, strain rate, and diastolic thickness [34]. In a multi-classification study of 163 patients (85 HCM, 30 AFD, 30 HHD, 18 CA), the 3D-MDA neural network model achieved an AUC of 0.94 (92% sensitivity, 90% specificity) in five-fold cross-validation, far superior to traditional threshold-based methods (regional wall thickness alone: AUC=0.75), demonstrating that combined structural and deformation analysis adds diagnostic information.

In deep learning approaches, Weberling et al. enrolled 400 patients from multiple centers and vendors (95 AL amyloidosis, 116 ATTR, 94 HCM, 95 healthy controls) from 56 medical institutions [35]. Employing a three-stage cascaded AutoML framework, the first stage distinguished healthy from patients with an AUC of 1.0; the second stage differentiated HCM from amyloidosis with an AUC of 0.99; and the third stage distinguished AL from ATTR with an AUC of 0.92. Agibetov et al. analyzed LGE, T1-mapping, and Cine sequences in 502 patients using VGG16 transfer learning to diagnose CA [36]. The study systematically compared the performance of three learning strategies across different sequences. LGE achieved the highest AUC of 0.96 with 94% sensitivity and 90% specificity, MOLLI sequence achieved an AUC of 0.93, and Cine sequence only 0.90, although multi-sequence fusion offers complementary value. Paciorek et al. employed DenseNet-161 architecture to develop a binary classification model on 200 patients (137 pathological, 63 healthy controls), comparing diagnostic performance between T1-mapping and LGE PSIR sequences [37]. The PSIR model achieved 88% accuracy and 100% sensitivity (38% specificity), while the T1-mapping model achieved an accuracy of 70%, a sensitivity of 78%, and a specificity of 54%.

With the evolution of deep learning architectures, multi-sequence fusion has progressively improved. Cockrum et al. employed a spatial-temporal Vision Transformer (ViT) model for three-class classification tasks in 807 patients (252 CA, 290 HCM, 265 others), integrating Cine and LGE sequences for differential diagnosis between CA and HCM [38]. The model achieved 84.1% accuracy and 0.954 AUC in the internal test set, maintaining 82.8% accuracy and 0.957 AUC performance in external validation cohorts. Notably, the ViT model maintained 61.1% accuracy on cases with low physician diagnostic confidence or human errors. After exclusion of image quality, the accuracy in this subgroup increased to 79.1%, demonstrating that advanced deep learning models based on multi-sequence fusion hold promise as valuable clinical decision support tools, particularly providing critical assistance in diagnostically challenging cases.

To systematically test the effectiveness of multi-sequence fusion, Wang et al. developed a two-stage diagnostic paradigm in a large-scale study of 9,719 patients (8,066 cardiovascular disease patients with 11 disease categories, 1,653 healthy controls) [39]. The first stage used Cine sequences for abnormality screening, while the second stage integrated Cine and LGE for specific disease classification. Multi-sequence fusion improved performance over single modality: class-weighted AUC increased by 1.9% and F1 score by 6.8% compared to short-axis Cine.

In cardiac amyloidosis subtype differentiation, Germain et al. applied VGG16 models to classify Cine and LGE images of 120 patients (70 AL, 50 ATTR) [40]. Cine-based model achieved 75.0% accuracy and AUC 0.839, outperforming LGE-based models (accuracy =61.1%, AUC=0.679) and CNN models (accuracy =84.5%, AUC=0.752), exceeding performance of three experienced physicians. Summaries of representative studies are provided in **Table 2**.

3 AI-BASED DIAGNOSIS OF CARDIAC HYPERTROPHY FROM MRI REPORTS

CMR reports represent the final textual output of clinical diagnostic decision-making, integrating expert interpretations of imaging findings with key quantitative measurements. The application of Natural Language Processing (NLP) techniques to automatically extract structured information from these unstructured text reports provides an efficient pathway for large-scale case screening, phenotypic classification, and clinical research.

Dewaswala et al. conducted pioneering work to develop an NLP system capable of automatically and precisely extracting HCM diagnoses along with phenotypic features from free-text CMR reports [41]. The study used a rule-based two-tier NLP method. In the first tier, the determined whether a report contained an HCM diagnosis. In the second tier, 9 key categorical concepts and 5 numerical concepts were extracted from the diagnosis-positive reports. Using a dataset of 391 CMR reports, the system demonstrated excellent performance, achieving an accuracy of 0.99 in extracting HCM diagnosis (sensitivity of 0.98 and specificity of 1.00). For phenotypic feature extraction, most categorical concepts achieved accuracies between 0.92 and 0.99, while numerical variables (such as LVEF and LV mass) achieved even higher accuracies of 0.96 to 0.99.

Sundaram et al. further explored a machine learning approach for automatically detecting HCM using CMR reports [42]. In their study, NLP techniques were applied to extract features from the “Impression” section of CMR reports and train an interpretable machine learning model to classify patients as “HCM-Yes”, “HCM-No”, or “HCM-Possible”. Researchers first applied unsupervised NLP to extract high-frequency co-occurring noun-adjective phrases from the text as candidate features, followed by training and predicting with random forest classifiers. In a data set with 1,835 reports, the resulting

Table 2. AI-assisted diagnosis of cardiac hypertrophy based on multi-sequence CMR fusion

Ref.	AI method	Sample size	Cardiac MRI sequence	Sensitivity (%)	Specificity (%)	Accuracy (%)	AUC
Zhang et al. [31]	ML	HCM: 421, HHD: 200	Cine, LGE	95.2	88.3	95.3	0.979
Kong et al. [32]	ML	HCM: 131, HHD: 48, NOR: 40	Cine, LGE, Native T1 mapping	-	-	-	0.906
Satriano et al. [34]	ML	HCM: 85, HTNcm: 30, AFD: 30, CA: 18	Cine, LGE	92.0	90.0	91.0	0.940
Weberling et al. [35]	ML	HCM: 94, AL CA: 95, ATTR CA: 116, NOR: 95	Cine, LGE, Native T1 mapping, T2 mapping	96.0	96.0	-	1.000
Agibetov et al. [36]	Fine-tuned VGGNet	CA: 82, other: 420	LGE, MOLLI (T1 mapping), CINE	94.0	90.0	-	0.960
Paciorek et al. [37]	DenseNet-161	HCM, CA, AL, NOR	T1 mapping, LGE PSIR	100.0	38.0	88.0	0.839
Cockrum et al. [38]	ViT	CA: 303, HCM: 339, Other: 322	Cine, LGE	-	-	84.1	0.945
Wang et al. [39]	VST	HCM: 2327, HHD: 402, NOR: 1250	Cine, LGE	100.0	98.6	99.1	0.991
Germain et al. [40]	VGG16	AL: 70, ATTR: 50	Cine, LGE	-	-	75.0	0.839

Note: AFD, Anderson–Fabry disease; AL, light chains; ATTR, transthyretin; AUC, area under the curve; CA, cardiac amyloidosis; CNN, convolutional neural networks; DL, deep learning; DCM, dilated cardiomyopathy; EMF, endomyocardial fibrosis; HCM, hypertrophic cardiomyopathy; HHD, hypertensive heart disease; HTNcm, hypertensive cardiomyopathy; LVH, left-ventricular hypertrophy; ML, machine learning; NOR, normal person; ViT, Vision Transformer; VST, Video Swin Transformer.

models achieved an overall classification accuracy of approximately 85%.

To conclude, AI-assisted diagnostic research based on CMR reports demonstrates that NLP and ML can accurately identify cardiac hypertrophy from unstructured report text, particularly in the automatic detection and phenotypic stratification of HCM. However, text-based models remain constrained by variability in report-writing styles and potential class imbalance within datasets, with limited capability for subtle differentiation between HCM and other hypertrophy types. Future work should consider integrating report-level NLP with multi-sequence imaging data to establish unified AI diagnostic system that combine textual and imaging information, thereby improving model generalizability and facilitating clinical translation.

4 CONCLUSION

This review systemically summarizes recent advances in the application of AI in diagnosing cardiac hypertrophy based on CMR imaging. It provides a comprehensive overview of AI diagnostic methods utilizing Cine-MRI sequences, T1/T2 mapping sequences, LGE sequences, and various multi-sequence fusion strategies. In addition, the technological development from traditional machine learning to deep learning is discussed, along with their applications in distinguishing normal from hypertrophic hearts and in the etiological classification of cardiac hypertrophy. Furthermore, this review highlights the emerging role of NLP-based automatic parsing of CMR reports in facilitating large-scale case identification and phenotypic stratification.

Despite these advances, several challenges remain. Data quality and annotation problems are major constraints to AI model performance. Most current studies are based on single-center datasets with relatively small sample sizes and potential class imbalances, thereby limiting generalizability of the developed models. Additionally, further development of AI technology still requires richer data support, including fusion of multimodal data and integration of clinical data. Combining information from genomics, proteomics, and other fields may enable a more comprehensive understanding of underlying mechanisms of cardiac hypertrophy, thereby enhancing the diagnostic and predictive accuracy of AI models. Moreover, with the further development of AI-based intelligent diagnostic technology, how to realize multi-center data sharing and model cooperation under the premise of protecting patients' privacy is also an important problem for the promotion of AI clinical application.

Looking forward, the application of AI in CMR-based diagnostic assistance for cardiac hypertrophy is promising. With the continuous evolution of deep learning algorithms and increasing computational capacity, AI models are expected to achieve higher levels of diagnostic accuracy and predictive capability. Strengthened interdisciplinary collaboration among medical imaging specialists, AI researchers and clinicians will be essential to guarantee that technological developments effectively address real clinical needs. With ongoing development in technology and accelerating clinical implementation of AI, these approaches are expected to play an increasingly important role in the precision diagnosis and individualized management of cardiac hypertrophy, thereby improving patient outcomes and quality of life.

DECLARATIONS

Author contributions

Shimin Zhou was responsible for the conception, drafting, revision, and summarization of the entire manuscript. Xudong Guo, Yunli Shen, Qinfen Jiang, Xin Gong, Jie Ding, Yihong Yang were in charge of the conception and revision of the manuscript, providing guiding support, and notifying the status of the manuscript. Guojie Xu was tasked with the acquisition, analysis, and integration of literature pertaining to AI diagnosis based on Cine-MRI sequence. Jican Wen was responsible for the acquisition, analysis, and integration of literature related to AI diagnosis based on T1/T2 Mapping and LGE sequences. Jingyang Niu was in charge of the acquisition, analysis, and integration of literature concerning AI diagnosis based on multi-sequence CMR data fusion and NLP-based report analysis.

Funding

This work is supported by Open Project Fund of State Key Laboratory of Cardiovascular Diseases (2024SKL-TJ012).

Data availability

Not applicable.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

Not applicable.

REFERENCES

[1] Tore D, Faletti R, Gaetani C, Bozzo E, Biondo A, Carisio A, et al. Cardiac magnetic resonance of hypertrophic heart phenotype: A review. Heliyon. 2023 Jun;9(6):e17336. https://doi.org/10.1016/j.heliyon.2023.e17336

[2] Cirillo C, Matarrese MAG, Monda E, Pagnano ME, Vitale J, Verrillo F, et al. Artificial intelligence for left ventricular hypertrophy detection and differentiation on echocardiography, cardiac magnetic resonance and cardiac computed tomography: A systematic review. Int J Cardiol. 2025 Mar 1;422:132979. https://doi.org/10.1016/j.ijcard.2025.132979

[3] Baggiano A, Del Torto A, Guglielmo M, Muscogiuri G, Fusini L, Babbaro M, et al. Role of CMR mapping techniques in cardiac hypertrophic phenotype. diagnostics (Basel). 2020 Sep 29;10(10):770. https://doi.org/10.3390/diagnostics10100770

[4] Lin A, Kolossváry M, Išgum I, Maurovich-Horvat P, Slomka PJ, Dey D. Artificial intelligence: Improving the efficiency of cardiovascular imaging. Expert Rev Med Devices. 2020 Jun;17(6):565-577. https://doi.org/10.1080/17434440.2020.1777855

[5] Cau R, Pisu F, Suri JS, Montisci R, Gatti M, Mannelli L, et al. Artificial intelligence in the differential diagnosis of cardiomyopathy phenotypes. Diagnostics (Basel). 2024 Jan 10;14(2):156. https://doi.org/10.3390/diagnostics14020156

[6] Budai A, Suhai FI, Csorba K, Dohy Z, Szabo L, Merkely B, et al. Automated classification of left ventricular hypertrophy on cardiac MRI. Appl Sci. 2022 Apr 20;12(9):4151. https://doi.org/10.3390/app12094151

[7] You Y, Viktorovich LA, Qiu J, Nikolaevich KA, Vladimirovich BY. Cardiac magnetic resonance image diagnosis of hypertrophic obstructive cardiomyopathy based on a double-branch neural network. Comput Methods Programs Biomed. 2021 Mar;200:105889. https://doi.org/10.1016/j.cmpb.2020.105889

[8] Ma ZP, Wang SW, Xue LY, Zhang XD, Zheng W, Zhao YX, et al. A study on the application of radiomics based on cardiac MR non-enhanced cine sequence in the early diagnosis of hypertensive heart disease. BMC Med Imaging. 2024 May 27;24(1):124. https://doi.org/10.1186/s12880-024-01301-9

[9] Abramikas Ž, Jasiukevičiūtė I, Balčiūnaitė G, Glaveckaitė S, Palionis D, Valevičienė N. Artificial intelligence performance in cardiac magnetic resonance strain analysis for aortic stenosis: Validation with echocardiography and healthy controls. Medicina (Kaunas). 2025 May 22;61(6):950. https://doi.org/10.3390/medicina61060950

[10] Zhang X, Cui C, Zhao S, Xie L, Tian Y. Cardiac magnetic resonance radiomics for disease classification. Eur Radiol. 2023 Apr;33(4):2312-2323. https://doi.org/10.1007/s00330-022-09236-x

[11] Germain P, Vardazaryan A, Padoy N, Labani A, Roy C, Schindler TH, et al. Classification of cardiomyopathies from MR cine images using convolutional neural network with transfer learning. Diagnostics (Basel). 2021 Aug 27;11(9):1554. https://doi.org/10.3390/diagnostics11091554

[12] Guo J, Lu H, Chen Y, Zeng M, Jin H. Artificial intelligence study on left ventricular function among normal individuals, hypertrophic cardiomyopathy and dilated cardiomyopathy patients using 1.5T cardiac cine MR images obtained by SSFP sequence. Br J Radiol. 2022 May 1;95(1133):20201060. https://doi.org/10.1259/bjr.20201060

[13] Wibowo A, Triadyaksa P, Sugiharto A, Sarwoko EA, Nugroho FA, Arai H, et al. Cardiac disease classification using two-dimensional thickness and few-shot learning based on magnetic

resonance imaging image segmentation. J Imaging. 2022 Jul 11;8(7):194. https://doi.org/10.3390/jimaging8070194

[14] Jiang S, Zhang L, Wang J, Li X, Hu S, Fu Y, et al. Differentiating between cardiac amyloidosis and hypertrophic cardiomyopathy on non-contrast cine-magnetic resonance images using machine learning-based radiomics. Front Cardiovasc Med. 2022 Oct 26;9:1001269. https://doi.org/10.3389/fcvm.2022.1001269

[15] Liu Q, Lu Q, Chai Y, Tao Z, Wu Q, Jiang M, et al. Papillary-muscle-derived radiomic features for hypertrophic cardiomyopathy versus hypertensive heart disease classification. Diagnostics (Basel). 2023 Apr 25;13(9):1544. https://doi.org/10.3390/diagnostics13091544

[16] Chuah SH, Md Sari NA, Chew BT, Tan LK, Chiam YK, Chan BT, et al. Phenotyping of hypertensive heart disease and hypertrophic cardiomyopathy using personalized 3D modelling and cardiac cine MRI. Phys Med. 2020 Oct;78:137-149. https://doi.org/10.1016/j.ejmp.2020.08.022

[17] Eckstein J, Moghadasi N, Körperich H, Weise Valdés E, Sciacca V, Paluszkiewicz L, et al. A machine learning challenge: Detection of cardiac amyloidosis based on bi-atrial and right ventricular strain and cardiac function. Diagnostics (Basel). 2022 Nov 4;12(11):2693. https://doi.org/10.3390/diagnostics12112693

[18] Diao K, Liang HQ, Yin HK, Yuan MJ, Gu M, Yu PX, et al. Multi-channel deep learning model-based myocardial spatial-temporal morphology feature on cardiac MRI cine images diagnoses the cause of LVH. Insights Imaging. 2023 Apr 24;14(1):70. https://doi.org/10.1186/s13244-023-01401-0

[19] Chen WW, Kuo L, Lin YX, Yu WC, Tseng CC, Lin YJ, et al. A deep learning approach to classify fabry cardiomyopathy from hypertrophic cardiomyopathy using cine imaging on cardiac magnetic resonance. Int J Biomed Imaging. 2024 Apr 26; 2024:6114826. https://doi.org/10.1155/2024/6114826

[20] Germain P, Vardazaryan A, Padoy N, Labani A, Roy C, Schindler TH, et al. Deep learning supplants visual analysis by experienced operators for the diagnosis of cardiac amyloidosis by cine-CMR. Diagnostics (Basel). 2021 Dec 29;12(1):69. https://doi.org/10.3390/diagnostics12010069

[21] Zhou H, Li L, Liu Z, Zhao K, Chen X, Lu M, et al. Deep learning algorithm to improve hypertrophic cardiomyopathy mutation prediction using cardiac cine images. Eur Radiol. 2021 Jun;31(6):3931-3940. https://doi.org/10.1007/s00330-020-07454-9

[22] Ferreira VM, Piechnik SK. CMR parametric mapping as a tool for myocardial tissue characterization. Korean Circ J. 2020 Aug;50(8):658-676. https://doi.org/10.4070/kcj.2020.0157

[23] Muser D, Chahal AA, Selvanayagam JB, Nucifora G. Clinical applications of cardiac magnetic resonance parametric mapping. Diagnostics (Basel). 2024 Aug 20;14(16):1816. https://doi.org/10.3390/diagnostics14161816

[24] Antonopoulos AS, Boutsikou M, Simantiris S, Angelopoulos A, Lazaros G, Panagiotopoulos I, et al. Machine learning of native T1 mapping radiomics for classification of hypertrophic cardiomyopathy phenotypes. Sci Rep. 2021 Dec 8;11(1):23596. https://doi.org/10.1038/s41598-021-02971-z

[25] Shi RY, Wu R, An DL, Chen BH, Wu CW, Du L, et al. Texture analysis applied in T1 maps and extracellular volume obtained using cardiac MRI in the diagnosis of hypertrophic cardiomyopathy and hypertensive heart disease compared with normal controls. Clin Radiol. 2021 Mar;76(3):236.e239-236.e219. https://doi.org/10.1016/j.crad.2020.11.001

[26] Huang S, Shi K, Zhang Y, Yan WF, Guo YK, Li Y, et al. Texture analysis of T2-weighted cardiovascular magnetic resonance imaging to discriminate between cardiac amyloidosis and hypertrophic cardiomyopathy. BMC Cardiovasc Disord. 2022 May 21; 22(1):235. https://doi.org/10.1186/s12872-022-02671-0

[27] Hwang IC, Chun EJ, Kim PK, Kim M, Park J, Choi HM, et al. Automated extracellular volume fraction measurement for diagnosis and prognostication in patients with light-chain cardiac amyloidosis. PLoS One. 2025 Jan 22;20(1):e0317741. https://doi.org/10.1371/journal.pone.0317741

[28] Burrage MK, Ferreira VM. Cardiovascular magnetic resonance for the differentiation of left ventricular hypertrophy. Curr Heart Fail Rep. 2020 Oct;17(5):192-204. https://doi.org/10.1007/s11897-020-00481-z

[29] Martini N, Aimo A, Barison A, Della Latta D, Vergaro G, Aquaro GD, et al. Deep learning to diagnose cardiac amyloidosis from cardiovascular magnetic resonance. J Cardiovasc Magn Reson. 2020 Dec 7;22(1):84. https://doi.org/10.1186/s12968-020-00690-4

[30] Zhou XY, Tang CX, Guo YK, Tao XW, Chen WC, Guo JZ, et al. Diagnosis of cardiac amyloidosis using a radiomics approach applied to late gadolinium-enhanced cardiac magnetic resonance images: A retrospective, multicohort, diagnostic study. Front Cardiovasc Med. 2022 Mar 30;9:818957. https://doi.org/10.3389/fcvm.2022.818957

[31] Zhang H, Tian J, Zhang C, Wang H, Hui K, Wang T, et al. Discrimination models with radiomics features derived from cardiovascular magnetic resonance images for distinguishing hypertensive heart disease from hypertrophic cardiomyopathy. Cardiovasc Diagn Ther. 2024 Feb 15;14(1):129-142. https://doi.org/10.21037/cdt-23-350

[32] Kong LC, Wu LM, Wang Z, Liu C, He B. An integrated algorithm for differentiating hypertrophic cardiomyopathy from hypertensive heart disease. J Magn Reson Imaging. 2023 Oct; 58(4):1084-1097. https://doi.org/10.1002/jmri.28580

[33] Lu C, Wang YG, Zaman F, Wu X, Adhaduk M, Chang A, et al. Predicting adverse cardiac events in sarcoidosis: Deep learning from automated characterization of regional myocardial remodeling. Int J Cardiovasc Imaging. 2022 Aug;38(8):1825-1836. https://doi.org/10.1007/s10554-022-02564-5

[34] Satriano A, Afzal Y, Sarim Afzal M, Fatehi Hassanabad A, Wu C, Dykstra S, et al. Neural-network-based diagnosis using 3-dimensional myocardial architecture and deformation: Demonstration for the differentiation of hypertrophic cardiomyopathy. Front Cardiovasc Med. 2020 Nov 11;7:584727. https://doi.org/10.3389/fcvm.2020.584727

[35] Weberling LD, Ochs A, Benovoy M, Aus dem Siepen F, Salatzki J, Giannitsis E, et al. Machine learning to automati-

cally differentiate hypertrophic cardiomyopathy, cardiac light chain, and cardiac transthyretin amyloidosis: A multicenter CMR study. Circ Cardiovasc Imaging. 2025 Jul;18(7):e017761. https://doi.org/10.1161/circimaging.124.017761

[36] Agibetov A, Kammerlander A, Duca F, Nitsche C, Koschutnik M, Donà C, et al. Convolutional neural networks for fully automated diagnosis of cardiac amyloidosis by cardiac magnetic resonance imaging. J Pers Med. 2021 Dec 1;11(12):1268. https://doi.org/10.3390/jpm11121268

[37] Paciorek AM, von Schacky CE, Foreman SC, Gassert FG, Gassert FT, Kirschke JS, et al. Automated assessment of cardiac pathologies on cardiac MRI using T1-mapping and late gadolinium phase sensitive inversion recovery sequences with deep learning. BMC Med Imaging. 2024 Feb 13;24(1):43. https://doi.org/10.1186/s12880-024-01217-4

[38] Cockrum J, Nakashima M, Ammoury C, Rizkallah D, Mauch J, Lopez D, et al. Leveraging a vision transformer model to improve diagnostic accuracy of cardiac amyloidosis with cardiac magnetic resonance. JACC Cardiovasc Imaging. 2025 Mar;18(3):278-290. https://doi.org/10.1016/j.jcmg.2024.09.010

[39] Wang YJ, Yang K, Wen Y, Wang P, Hu Y, Lai Y, et al. Screening and diagnosis of cardiovascular disease using artificial intelligence-enabled cardiac magnetic resonance imaging. Nat Med. 2024 May;30(5):1471-1480. https://doi.org/10.1038/s41591-024-02971-2

[40] Germain P, Vardazaryan A, Labani A, Padoy N, Roy C, El Ghannudi S. Deep learning to classify AL versus ATTR cardiac amyloidosis MR images. Biomedicines. 2023 Jan 12;11(1):193. https://doi.org/10.3390/biomedicines11010193

[41] Dewaswala N, Chen D, Bhopalwala H, Kaggal VC, Murphy SP, Bos JM, et al. Natural language processing for identification of hypertrophic cardiomyopathy patients from cardiac magnetic resonance reports. BMC Med Inform Decis Mak. 2022 Oct 18;22(1):272. https://doi.org/10.1186/s12911-022-02017-y

[42] Sundaram DSB, Arunachalam SP, Damani DN, Farahani NZ, Enayati M, Pasupathy KS, et al. Natural language processing based machine learning model using cardiac MRI reports to identify hypertrophic cardiomyopathy patients. Proc Des Med Devices Conf. 2021 Apr;2021:V001T03A005. https://doi.org/10.1115/dmd2021-1076

Progress in
Medical Devices

LETTER TO THE EDITOR

Slim exquisite easy-exposing video laryngoscope: A novel video laryngoscope

Chenglong Zhu, Zui Zou

School of Anesthesiology, Naval Medical University, Shanghai 200433, China.

Corresponding author: Zui Zou.

Address correspondence to: Zui Zou, School of Anesthesiology, Naval Medical University, 800 Xiangyin Road, Yangpu District, Shanghai 200433, China.
E-mail: zouzui@smmu.edu.cn.

Received February 16, 2026; Accepted March 16, 2026; Published March 31, 2026

DOI: 10.61189/551629zyhfiv

With the evolution of traditional direct laryngoscopes into video-assisted laryngoscopes, the viewing angle provided by video laryngoscopes has been substantially enlarged, enabling more intuitive and clearer visualization of pharyngeal structures and the glottis. However, the design of video laryngoscopes generally retains the relatively bulky blade carrier of traditional direct laryngoscopes. During routine use, this design may still limit the field of view, necessitating significant jaw elevation to obtain a clear view [1, 2]. To address this limitation, our team developed a slim exquisite easy-exposing video laryngoscope (SEE-VL), a slender and refined device designed to facilitate easier glottic visualization (Registration Certificate No.: Su Xie Zhun 20252082044).

The most significant difference between SEE-VL and traditional video laryngoscopes (e.g., UESCOPE® video laryngoscope) lies in the optimized cross-sectional design of blade carrier. While ensuring adequate exposure of the laryngeal structures, SEE-VL minimizes additional trauma to the oral cavity and larynx, providing more intraoral space for establishing an artificial airway. Additionally, SEE-VL is equipped with a high-resolution display, enabling clearer visualization of the laryngeal structures and glottis (**Figure 1**).

Beyond routine airway establishment, the slim blade design of SEE-VL is particularly suitable for patients with anticipated difficult airways, including those with limited mouth opening, restricted head and neck mobility, or missing teeth—conditions commonly observed in patients with maxillofacial trauma, temporomandibular joint disorders, cervical spine surgery, or obesity). This novel SEE-LV may broaden the clinical applicability of video laryngoscopy in challenging airway scenarios.

DECLARATIONS

Author contributions

Zui Zou conceived the study, and Chenglong Zhu draft the manuscript.

Funding

This research received no external funding.

Data availability

Not applicable.

Ethics approval and consent to participate

Not applicable.

Consent for publication

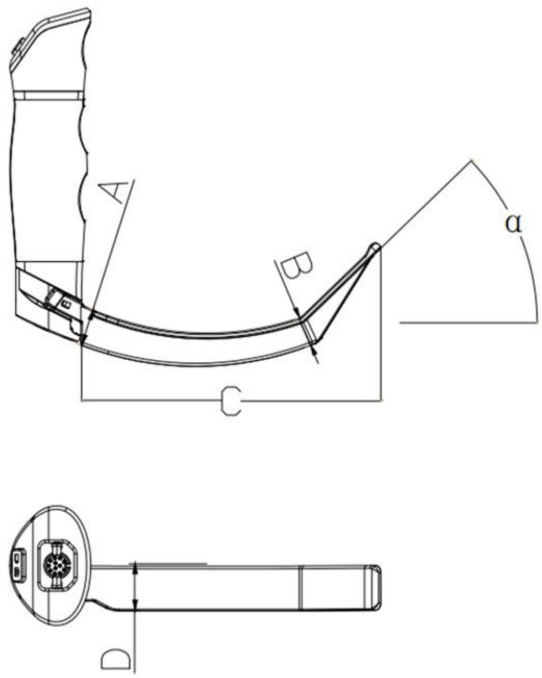
Not applicable.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

Not applicable.



Parameters	Size A (mm)	Size B (mm)	Size C (mm)	Size D (mm)	Angle α (°)
Laryngoscope blade (SEE-VL)	13.2	10.5	106.9	16.7	44.0
Laryngoscope blade (UE)	16.7	12.5	106.9	21.7	45.0

Figure 1. Schematic diagram of the bade carrier in the SEE-VL and its comparison with a conventional video laryngoscope (UESCOPE® video laryngoscope). SEE-VL, Slim Exquisite Easy-exposing Video Laryngoscope.

REFERENCES

[1] Sulser S, Ubmann D, Schlaepfer M, Brueesch M, Goliasch G, Seifert B, et al. C-MAC videolaryngoscope compared with direct laryngoscopy for rapid sequence intubation in an emergency department: A randomised clinical trial. Eur J Anaesthesiol. 2016 Dec;33(12):943-948. https://doi.org/10.1097/eja.0000000000000525

[2] Mazzinari G, Rovira L, Henao L, Ortega J, Casasempere A, Fernandez Y, et al. Effect of dynamic versus stylet-guided intubation on first-attempt success in difficult airways undergoing glidescope laryngoscopy: A randomized controlled trial. Anesth Analg. 2019 Jun;128(6):1264-1271. https://doi.org/10.1213/ane.0000000000004102

Progress in
Medical Devices

RESEARCH ARTICLE

A metallic foreign object detection algorithm in pharmaceuticals based on phase rotation and smoothed pseudo-Wigner-Ville distribution

Lin Jiang, Piding Li

Department of Health Sciences and Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China.

Corresponding author: Piding Li.

Address correspondence to: Piding Li, Department of Health Sciences and Engineering, University of Shanghai for Science and Technology, No. 334 Jungong Road, Yangpu District, Shanghai 200093, China.
E-mail: lipiding_usst@qq.com.

Received November 12, 2025; Accepted January 27, 2026; Published March 31, 2026

DOI: 10.61189/447159fjktza

Abstract

Currently, the pharmaceutical manufacturing industry faces problems such as low accuracy in detecting metal foreign objects due to product effects. To address this issue, this paper proposes a metal detection algorithm based on phase rotation and time-frequency analysis. Phase rotation suppresses product effect interference, and smoothed pseudo-Wigner-Ville distribution (SPWVD) is used to acquire time-frequency images, identifying significant differences representing metal foreign objects and achieving metal detection under strong product effect interference. To ensure computational efficiency meets industrial real-time requirements, an embedded software system with a dual-core CPU and CLA working in tandem is employed, improving algorithm efficiency through hardware improvements. Test results show that the system achieves a detection accuracy exceeding 98% for 0.8 mm ferromagnetic metals and 1.2 mm non-ferromagnetic metals, with a single detection cycle completed within 10 ms.

Keywords: Medicine, Metal detection, Time-frequency analysis

1 INTRODUCTION

Metal contaminants may be introduced during the pharmaceutical manufacturing process. These contaminants typically originate from various production stages, such as equipment damage, human error, and product packaging. Even with all preventative measures, it is difficult to completely eliminate these metal contaminants [1, 2]. Therefore, integrating metal foreign object detection into the pharmaceutical manufacturing process has become a mandatory procedure. Failure to do so could harm those purchasing the drugs and cause significant economic losses to pharmaceutical companies [3, 4].

Currently, among the metal foreign object detection devices used by many industrial companies, those based on the principle of electromagnetic induction are the most popular type. These metal detectors generate a constantly changing magnetic field. Metal foreign objects within this area generate a secondary magnetic field due to the eddy current effect, which interferes with and cancels out the original magnetic field. By detecting changes in the magnetic field, the presence of metal contaminants can be identified [5, 6]. However, metal detection in pharmaceuticals is often affected by the product itself. When water-containing or electrically saturated pharmaceuticals are placed in an alternating magnetic field, they generate product signals that are similar to those generated by metals. The amplitude of these product signals may mask weak metal signals, making it difficult to detect small metal particles during pharmaceutical metal detection [7]. To address this issue, this study proposes a metal detection method based on time-frequency analysis. This method utilizes a lock-in amplifier circuit to separate the real and imaginary parts of the measurement signal, and then employs a time-frequency analysis algorithm to process and analyze the signal in the time-frequency domain. This enables the detection of minute metal contaminants even under strong product effect interference.

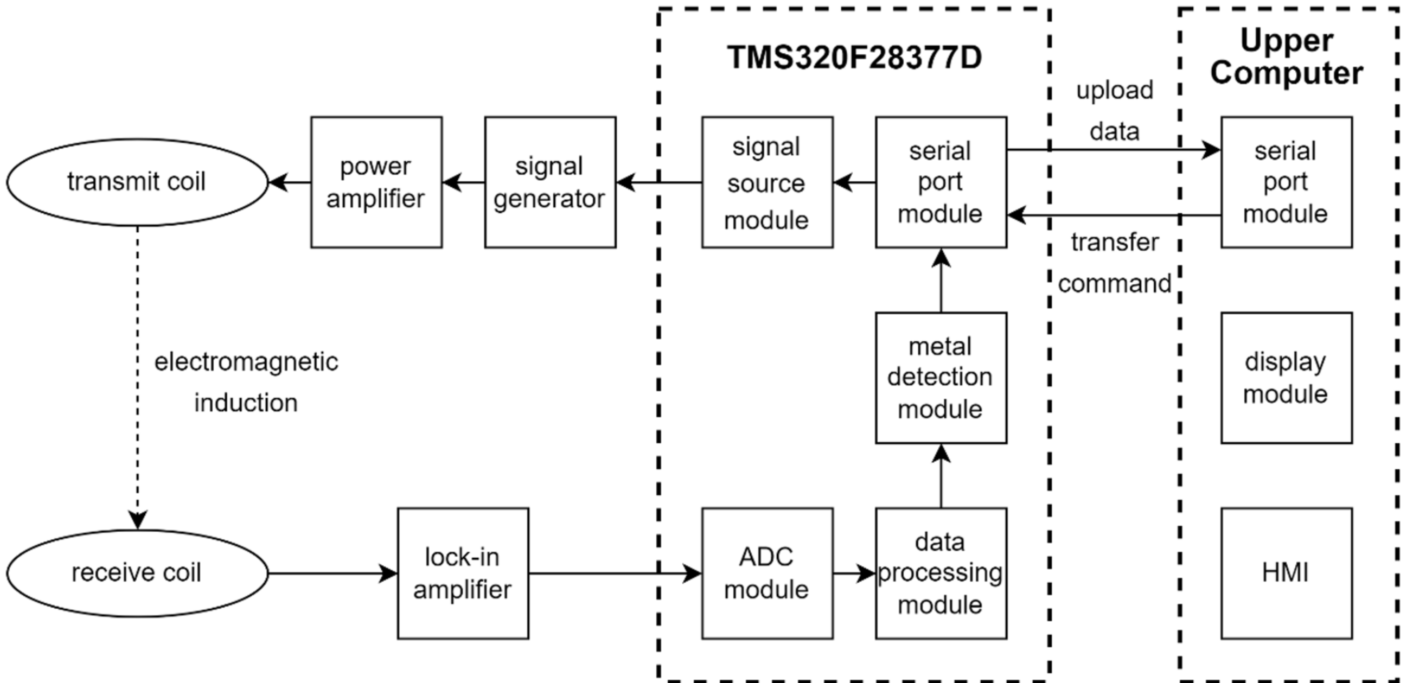


Figure 1. Metal detection system framework.

2 METAL DETECTION SYSTEM DESIGN

The overall structure of the metal detection system designed in this paper is shown in **Figure 1**, including a signal generator, a balanced coil sensor, a signal processing circuit, a main control chip TMS320F28377D, and a host computer.

The experimental method is as follows. Firstly, test samples were prepared, using an oral liquid with high product efficacy as the test drug. According to the national standard GB/T 25345-2010, the metal detector must be able to detect ferromagnetic materials with a diameter of 1.2 mm and non-ferromagnetic materials with a diameter of 1.5 mm. Therefore, iron samples with diameters of 1.2 mm, 1.0 mm, and 0.8 mm, and stainless steel and copper metal samples with diameters of 1.5 mm, 1.2 mm, and 1.0 mm were selected for the test. These metal samples were placed under the drug to simulate metal contaminants.

Next, the frequency and amplitude of the excitation signal were set via the host computer. The host computer transmits control commands to the main control chip via serial port. After decoding the commands, the main control chip controls the direct digital synthesizer to output corresponding sinusoidal excitation signals. These signals drive the excitation coil to generate corresponding alternating magnetic fields.

The system first needs to operate under no-load conditions without any samples passing through. Ideally, the balancing coil structure will cancel the induced electromotive force in the two receiving coils, making the system output approximately zero [8]. However, in practice, the two receiving coils of the system cannot be perfectly synchronized and need to be adjusted based on the output signal during no-load operation.

Subsequently, the test sample was allowed to pass through the system, breaking the original magnetic field balance, and the balancing coil will output a detection signal. After the detection signal was quadraturely demodulated by the lock-in amplifier, it would generate an in-phase signal (I) and a quadrature signal (Q). The main control chip synchronously acquires these two signals and uses an algorithm to detect metallic foreign objects. Finally, the detection results are uploaded to the host computer for display via serial communication.

3 METAL DETECTION ALGORITHM DESIGN

3.1 Signal preprocessing

In metal detection, electromagnetic interference introduces high-frequency noise into the signal, while the effective signal is typically concentrated in the low-frequency range. To suppress high-frequency noise while preserving the waveform characteristics of the effective signal, this study employs a Butterworth low-pass filter for preprocessing the original signal. This filter exhibits maximum flat amplitude characteristics within its passband. The specific design parameters of this filter are shown in **Table 1**.

Table 1. Low-pass filter parameters

Parameter	Sampling rate/Hz	Cutoff frequency/Hz	Order
Value	150	10	2

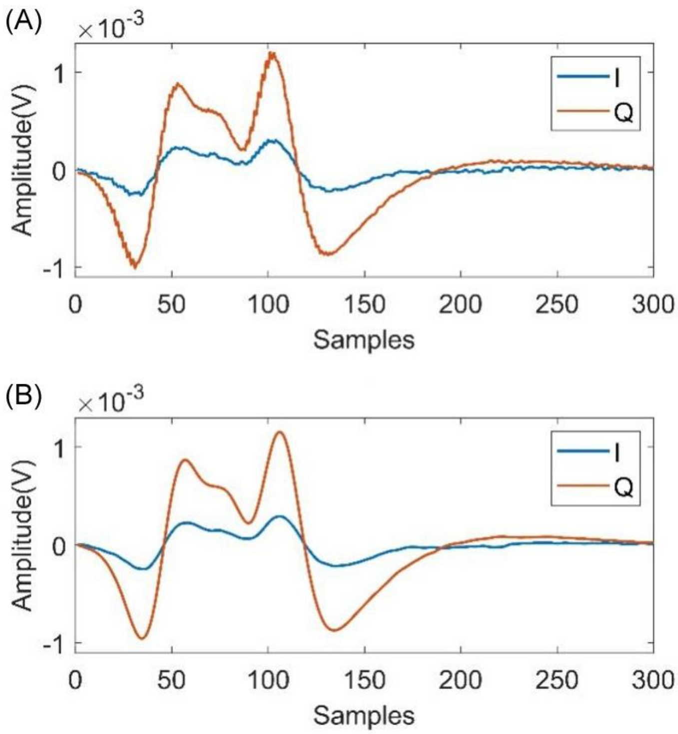


Figure 2. Signal before and after low-pass filter. (A) Raw signal; (B) Filtered signal.

Figure 2 shows a comparison of the time-domain waveforms of the signal before and after filtering, obtained using actual measurement data. The original signal shown in **Figure 2A** contains a large amount of superimposed high-frequency noise. The filtered signal shown in **Figure 2B** effectively suppresses these high-frequency noise components, resulting in a smoother waveform, while also preserving the main characteristics of the original signal.

3.2 Phase rotation

In metal detection of pharmaceuticals with high moisture content, interference from the "product effect" is frequently encountered. This means that the product itself generates a signal under the influence of eddy currents, and this product signal is typically much stronger than the metal signal, interfering with metal detection.

Figure 3A shows the IQ component diagrams of metals (iron (Fe), stainless steel (SUS), and brass (Nofe)), and **Figure 3B** displays the IQ component diagrams of metal-pharmaceutical mixtures. It can be observed that the phases of different metals and pharmaceutical materials differ. However, when a pharmaceutical is mixed with a metal, its phase is very close to that of the individual pharmaceutical material, making direct differentiation difficult. Therefore, this study employs a phase rotation algorithm to suppress interference from product effects by rotating the phase of the detection signal, thereby improving the ability to identify metal signals.

Figure 4 shows the IQ component plot of the pharmaceutical data and the fitted line obtained by the least squares method. Based on this fitted line, the characteristic phase of this type of pharmaceutical can be calculated. Then, this characteristic is used to rotate the detection signal. **Figure 5** presents a comparison of the signals before and after phase rotation.

Figure 5A shows that the Q component of the pharmaceutical detection signal is significantly suppressed after phase rotation. However, **Figure 5B** shows that the Q component of the detection signal containing the metal mixture still exhibits significant fluctuations after the same phase rotation.

3.3 Time-frequency analysis

Metal detection signals are typical non-stationary signals, with their frequency components changing dynamically over time. Traditional Fourier analysis can only analyze the global frequency domain information of the signal and cannot capture its time-varying spectral characteristics [9, 10]. This study introduces time-frequency analysis techniques, selecting three methods, the Short-Time Fourier Transform (STFT), Wigner-Ville Distribution (WVD), and Smooth Pseudo-Wigner-Ville Distribution (SPWVD)—to process the acquired actual metal detection signals. Through comparative analysis, the method most suitable for the application scenario of this study was selected. The time-frequency representations obtained by the three methods are shown in **Figure 6**.

STFT uses a sliding technique to divide the signal into windows to obtain the local time-frequency characteristics of the signal [11, 12]. However, this method has limited time-frequency resolution, making it difficult to meet the high-precision analysis requirements for transient and steady-state features [13]. As shown in **Figure 6A**, although the background of the STFT time-frequency representation is clean, its time-frequency resolution is the lowest, and the energy distribution is relatively dispersed, which is not conducive to the accurate extraction of subsequent subtle features.

WVD exhibits the best time-frequency clustering, but when the signal contains multiple frequency components, there will be spurious responses between these components without physical meaning, leading to severe cross-term interference [14-16]. As shown in **Figure 6B**, WVD time-frequency representation achieves the highest energy concentration, yet the

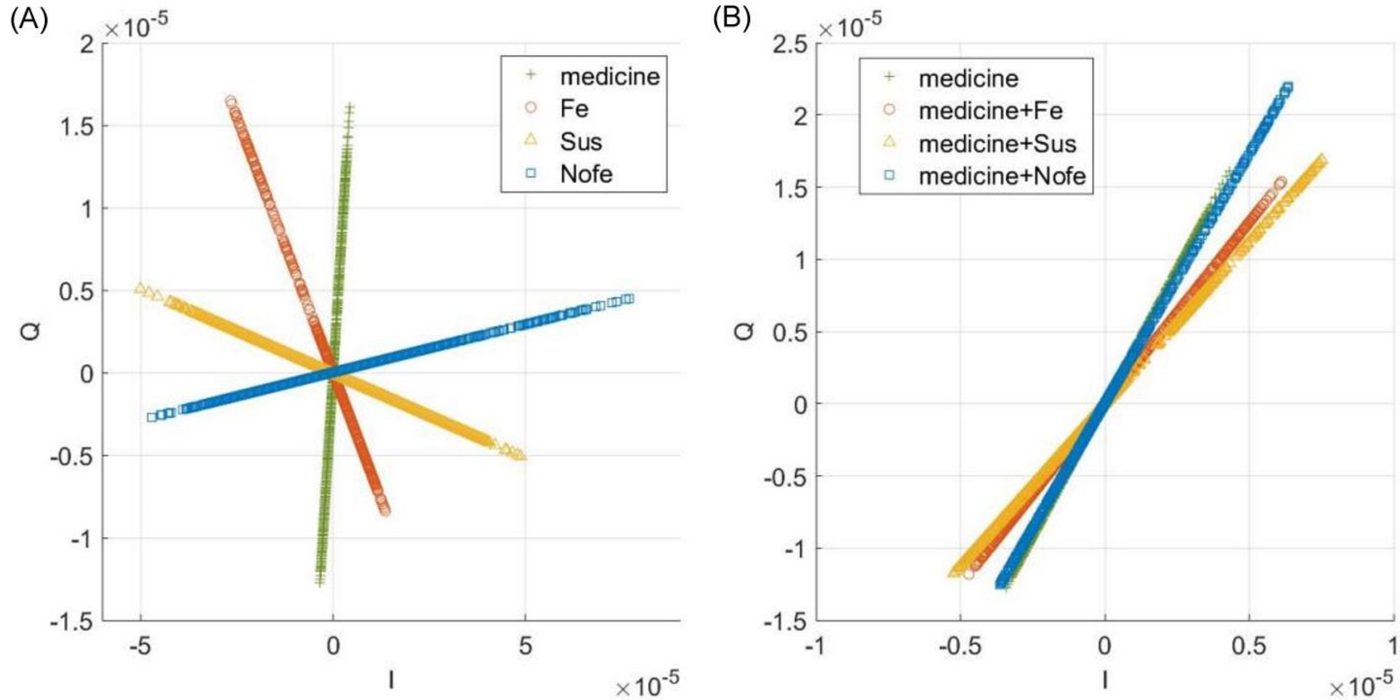


Figure 3. IQ component diagrams. (A) Metal and medicine; (B) Metal-pharmaceutical mixture.

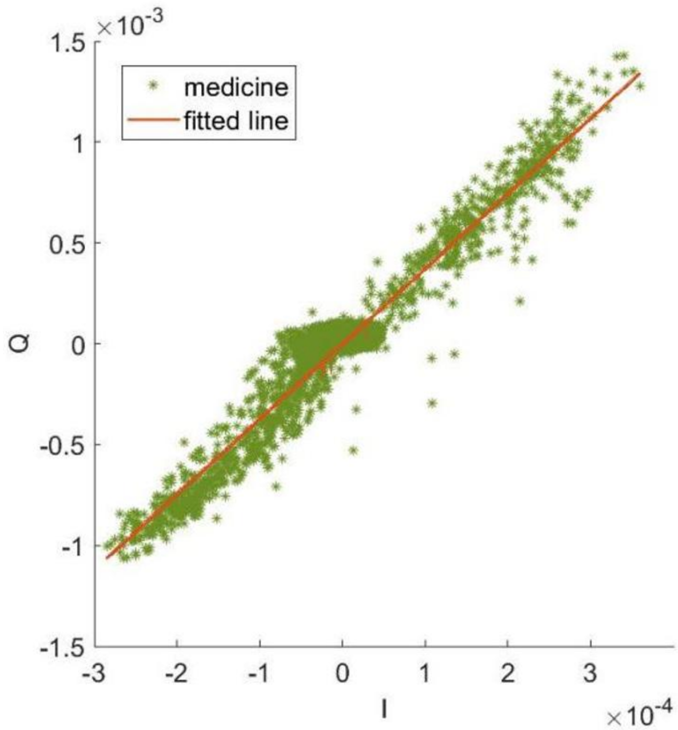


Figure 4. Phase fitting.

severe cross-term interference complicates the interpretation of genuine signal components, posing a significant challenge to subsequent classification tasks.

SPWVD, based on WVD, applies window functions for smoothing in both time and frequency dimensions [17, 18]. As shown in **Figure 6C**, SPWVD time-frequency representation achieves the best balance between time-frequency concentration and cross-term suppression. It inherits the energy concentration advantage of WVD while filtering out most cross-terms through a smoothing window, ultimately yielding a clear and accurate time-frequency representation [19].

Practical testing has proven that SPWVD is the most suitable method for analyzing non-stationary signals such as metal detection signals. Therefore, this study adopts SPWVD as the core algorithm for time-frequency analysis.

3.4 Metal discrimination algorithm

Product signals containing metal foreign objects (positive samples) and product signals without metal foreign objects (negative samples) exhibit stable and quantifiable response differences in the time-frequency domain. Based on this phenomenon, this study proposes a metal detection algorithm based on a time-frequency difference template. The metal detection algorithm consists of two parts: "region learning" and "classification and recognition". The algorithm flow is shown in **Figure 7**. In the "region learning" stage, five sets of positive samples and five sets of negative samples are collected respectively. The average time-frequency maps of the positive and negative samples are used to obtain the significantly different regions that highlight the metal features. In the "classification and recogni-

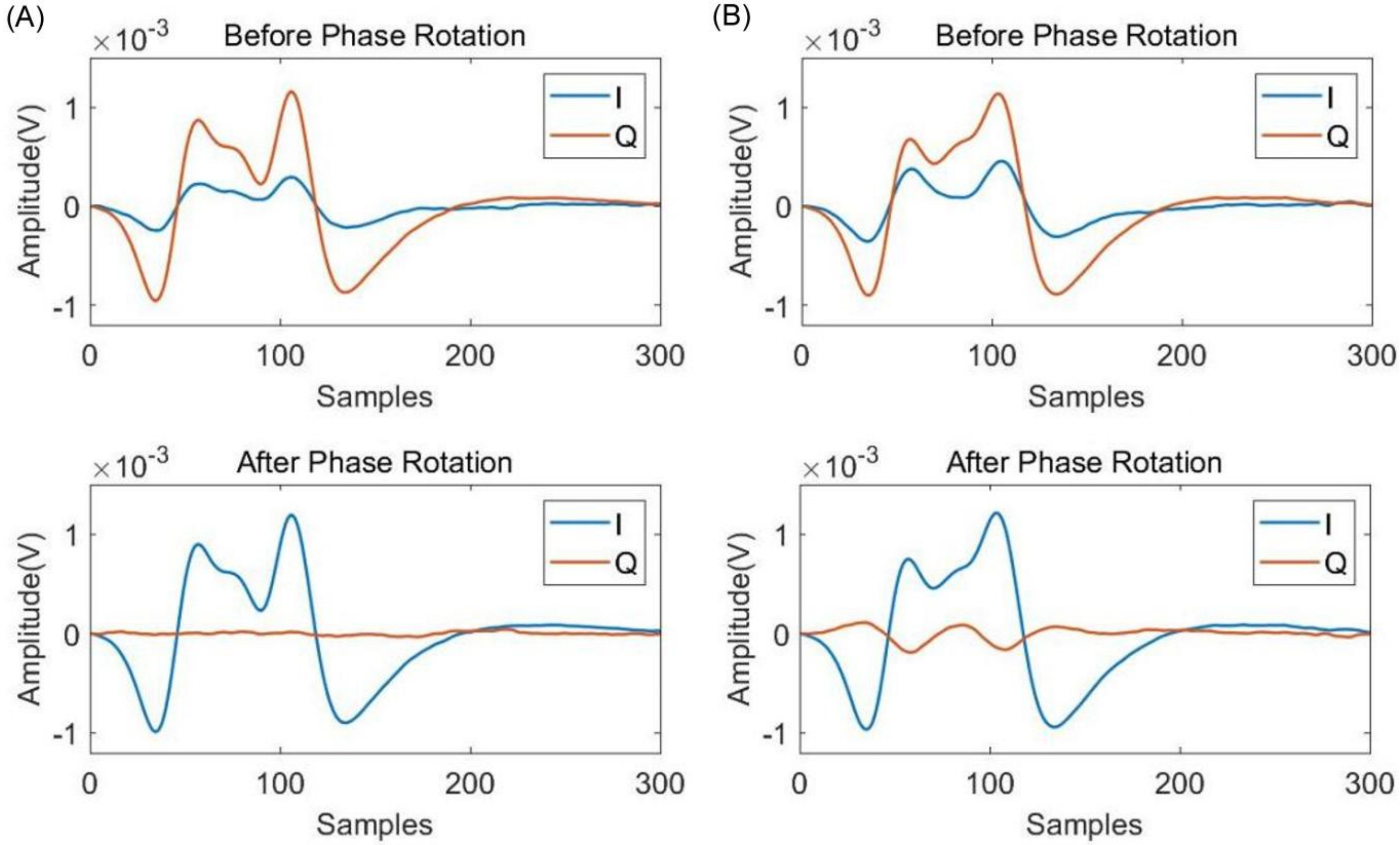


Figure 5. Signal before and after phase rotation. (A) Medicine; (B) Metal-pharmaceutical mixture.

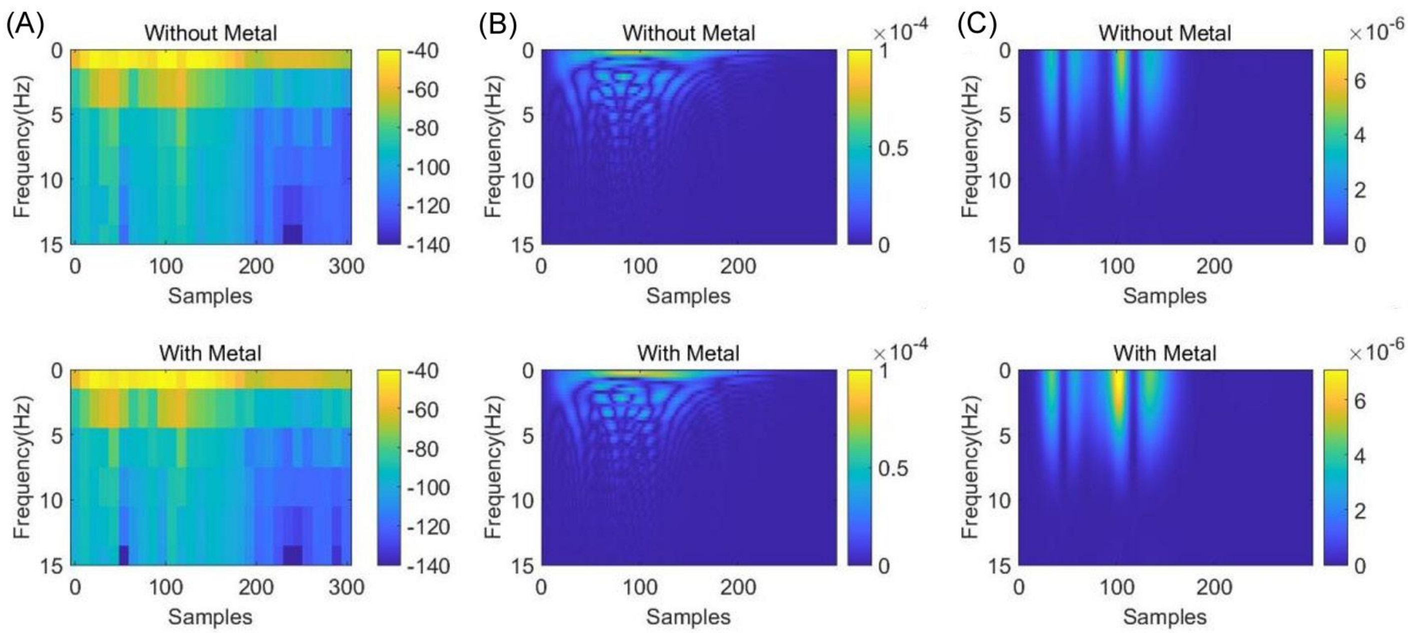


Figure 6. Time-frequency analysis spectrogram. (A) STFT spectrogram; (B) WVD spectrogram; (C) SPWVD spectrogram. STFT, Short-Time Fourier Transform; WVD, Wigner-Ville Distribution; SPWVD, Smoothed Pseudo and Smooth Pseudo-Wigner-Ville Distribution.

tion" stage, the response values of the test sample's time-frequency map and the average time-frequency map of the negative samples in the significantly different regions are compared, and then compared with a preset threshold to complete

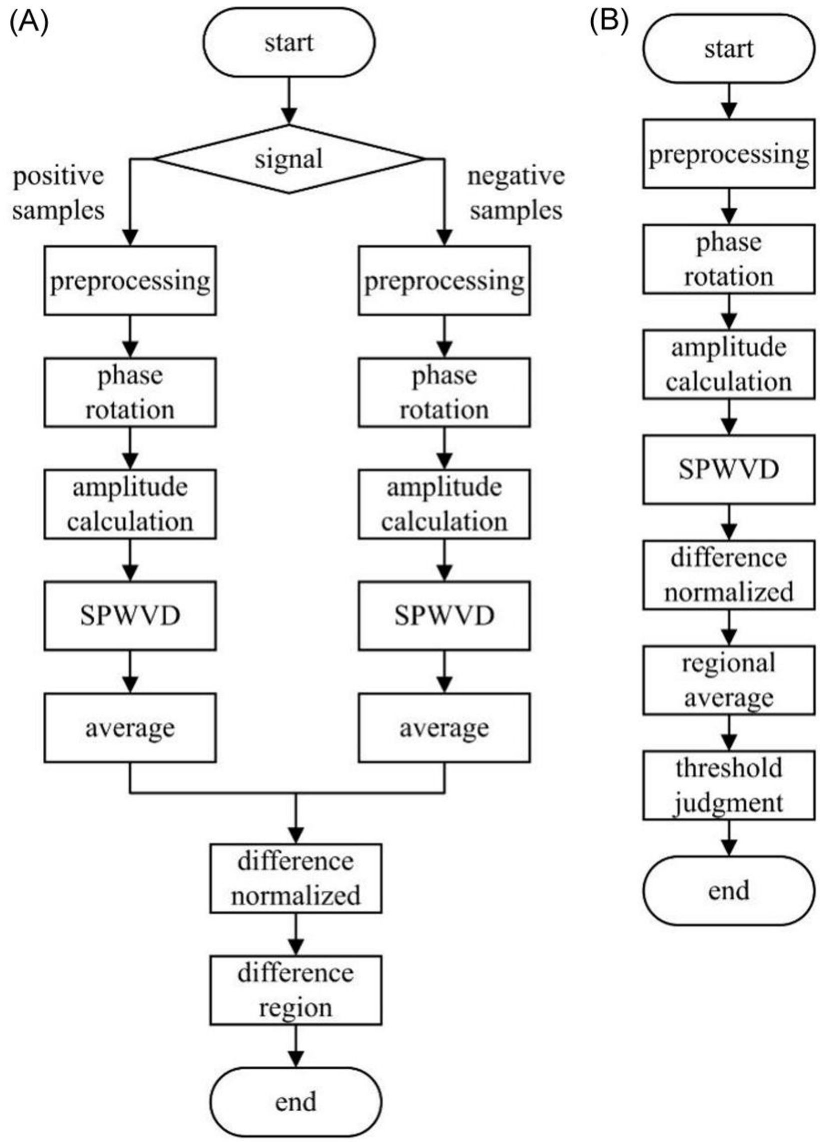


Figure 7. Metal discrimination flowchart. (A) Region learning; (B) Classification recognition.

Table 2. Computational time of algorithm modules

Module	Low-pass filtering	Amplitude	SPWVD	RFFT	Overall
Clock Cycles	737,093	19,524	2,236,384	1,616,905	3,043,692
Runtime/ms	3.685	0.098	11.182	8.085	15.218

the classification judgment, thereby realizing the detection of metal foreign objects.

4 EMBEDDED SYSTEM ALGORITHM DEPLOYMENT

The main control chip used in this system is the Texas Instruments TMS320F28377D microcontroller. This microcontroller integrates two CPUs with a clock frequency of 200 MHz and two programmable law accelerators (CLAs), which can perform computational tasks independently of the CPUs. Signal acquisition, data processing, and metal detection algorithms are all implemented on this microcontroller.

To optimize the system's detection performance, this study conducted actual tests on the computation time consumption of each module of the algorithm. The test results are shown in **Table 2**.

The test results indicate that the multiple real fast Fourier transforms (RFFTs) in the SPWVD algorithm are the most time-consuming part of the entire process. From the algorithm principle, the SPWVD calculation can be divided into two main steps: instantaneous autocorrelation and RFFT. Since multiple sets of instantaneous autocorrelation data are independent of each other, their RFFT calculations can be processed in parallel. Therefore, this study proposes a hardware acceleration scheme for the algorithm based on a dual-core CPU and CLA. The overall block diagram of this scheme is shown in **Figure 8**. A single set of detection data will be divided into two parts, which will be processed jointly by CPU1 and CPU2. The most time-consuming RFFT calculation task will be offloaded to the CLA for execution [20]. While the CLA is executing the RFFT calculation task for the current data block, the CPU can simultaneously execute the instantaneous autocorrelation calculation task for the next data block, thereby improving computational efficiency. Finally, CPU1 synthesizes the metal discrimination results to determine whether there is a metal foreign object.

The parallel processing scheme based on dual-core CPU and CLA involves data communication between the dual-core CPUs and between the CPU and the CLA. To ensure effective and coordinated data transmission, this paper proposes a data exchange and management scheme. CPU1 and CPU2 achieve data access synchronization through IPC, and achieve efficient data transmission between the dual-core CPUs using Global Shared RAM (GSRAM). The CPU and CLA achieve data access synchronization through Task Driver Management. This mechanism differs from the traditional interrupt-driven mode, does not involve interrupt overhead, and can achieve efficient data transmission between the CPU and the CLA through Local Shared RAM (LSRAM).

To test the performance of the optimization scheme, this study measured the running time of the detection algorithm under

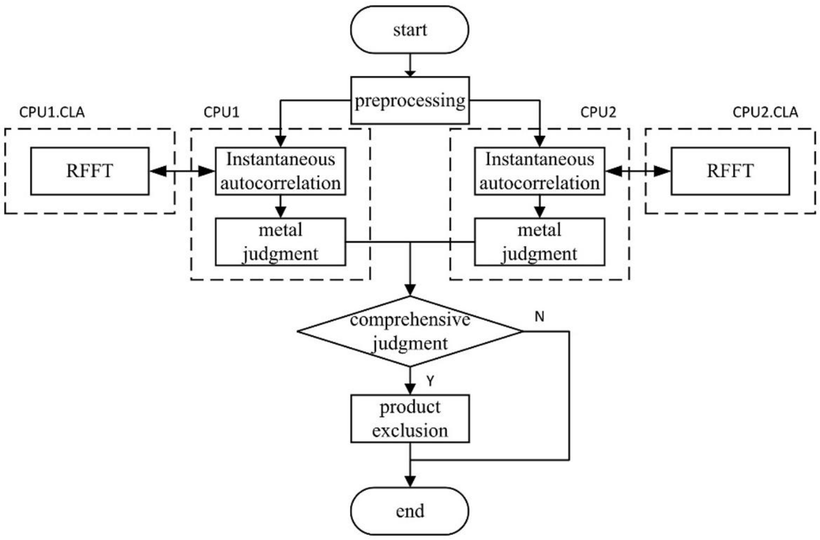


Figure 8. Hardware acceleration flowchart.

Table 3. Performance comparison of different implementation schemes

Scheme	Single-Core CPU	Single-Core CPU+CLA	Dual-Core CPU+CLA
Clock Cycles	3,043,692	2,493,186	1,894,567
Runtime/ms	15.218	12.466	9.473
Speedup Ratio/%	-	18.100	37.700

Table 4. Statistical results of metal detection

Contaminant	Dimensions/mm	Accuracy/%
None	-	100
Fe	1.200	100
	1.000	100
	0.800	98
SUS	1.500	100
	1.200	100
	1.000	86
Nofe	1.500	100
	1.200	100
	1.000	82

three different implementation methods through actual testing. The test results are shown in **Table 3** below. Test results show that compared to a single-core CPU solution, the combination of a single-core CPU and CLA reduces computation time by approximately 18.1%. Meanwhile, the dual-core CPU and CLA processing solution reduces time by approximately 37.7% compared to the single-core CPU solution. The dual-core CPU and CLA acceleration algorithm improve the system's real-time processing performance, and the reduced detection time allows for further increases in production line transmission speed, thereby improving production line efficiency.

5 RESULTS

This study selected an oral liquid with a strong product effect as the test product to verify the detection system's ability to detect different metals. The metal foreign objects used for testing were standard test sample cards placed inside plastic cards, containing three types of particles: iron, stainless steel, and brass. The test sample set consisted of positive samples containing some metal contaminants and negative samples without metals. Each sample was tested 50 times in the detection system, and the relevant statistical data are shown in **Table 4**.

The test results show that the system can stably detect ferromagnetic metals with a diameter of 1.2 mm and non-magnetic metals with a diameter of 1.5 mm, meeting the detection accuracy requirements of GB/T 25345-2010 for metal detectors. However, the detection accuracy of the system for metal-containing samples gradually decreases as the metal size decreases. For metal samples of the same size but different materials, the detection accuracy of ferromagnetic metals is much higher than that of non-ferromagnetic metals. Compared with other detection systems in other studies, this system has the same detection accuracy for ferromagnetic metals and can stably detect ferromagnetic metals as small as 0.8 mm, while the detection sensitivity for non-ferromagnetic metals has been improved to 1.0 mm, which significantly improves the detection performance for non-ferromagnetic metals [21].

6 CONCLUSION

This study implements an embedded metal detection system using a time-frequency analysis algorithm. Under strong product effect scenarios, the system can stably detect ferromagnetic metals of 0.8 mm and non-ferromagnetic metals of 1.2 mm. At the same time, a collaborative task processing structure based on a dual-core CPU and CLA was designed, which shortens the single-frame data processing time to about 37%, and controls the detection time to within 10 ms, meeting the requirements of industrial real-time detection.

DECLARATIONS

Author contributions

Lin Jiang is responsible for software development, data analysis, and writing the original draft, Piding Li provides supervision and participates in the review and editing of the manuscript.

Funding

This research received no external funding.

Data availability

Not applicable.

Ethics approval and consent to participate

Not applicable.

Consent for publication

All authors have reviewed the final version of the manuscript and have given their consent for publication.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

Not applicable.

REFERENCES

[1] Poonam K. Metallic impurities in pharmaceuticals: An overview. Curr Pharm Anal. 2021;17(8):960-968. http://dx.doi.org/10.2174/1573412916999200711151147

[2] Balaram V. Recent advances in the determination of elemental impurities in pharmaceuticals – Status, challenges and moving frontiers. Trends Anal Chem. 2016 Jun 1;80:83-95. https://doi.org/10.1016/j.trac.2016.02.001

[3] Abernethy DR, DeStefano AJ, Cecil TL, Zaidi K, Williams RL; USP Metal Impurities Advisory Panel. Metal impurities in food and drugs. Pharm Res. 2010 May 1;27(5):750-755. https://doi.org/10.1007/s11095-010-0080-3

[4] Aleluia ACM, Nascimento MdS, dos Santos AMP, dos Santos WNL, de Freitas Santos Júnior A, Ferreira SLC. Analytical approach of elemental impurities in pharmaceutical products: A worldwide review. Spectrochim Acta B. 2023 Jul 1;205:106689. https://doi.org/10.1016/j.sab.2023.106689

[5] Shi H, Xie Y, Zhang H, Li W, Ma L, Xu Z, et al. An ultrasensitive debris microsensor for oil health monitoring based on resistance–inductance parameter. IEEE Trans Instrum Meas. 2021 August 30;70:1-9. https://doi.org/10.1109/TIM.2021.3105255

[6] Xie Y, Shi H, Zhang H. Mixed metal differentiation method using microfluidic oil detection sensors. J Phys Conf Ser. 2024 Apr 1;2740:012014. https://doi.org/10.1088/1742-6596/2740/1/012014

[7] Haimovich H, Marelli D, Sarlinga D. A signal processing method for metal detection sensitivity improvement in balance-coil metal detectors for food products. 2020 IEEE International Conference on Industrial Technology (ICIT); 2020 Feb 26-28; Buenos Aires (Argentina). Piscataway (NJ): IEEE; 2020. p. 645-651. https://doi.org/10.1109/ICIT45562.2020.9067312

[8] Okabe S, Sasada I. Development of a planar-type high sensitivity metallic contaminant detector. Aip Advances. 2017 May; 7(5):7. https://doi.org/10.1063/1.4978590

[9] Wang Y, Ji S, Xu H. Non-stationary signals processing based on STFT. 2007 8th International Conference on Electronic Measurement and Instruments; 2007 Aug 16-18; Xi'an (China). Piscataway (NJ): IEEE; 2007. p. 3-301-3-304. https://doi.org/10.1109/ICEMI.2007.4350914

[10] Liu W, Ma Z, Sun H. Summary of weak signal detection and processing methods. Proceedings of the 2018 International Conference on Mechanical, Electronic, Control and Automation Engineering (MECAE 2018); 2018 Mar 30-31; Qingdao (China). Atlantis Press; 2018. p. 213-216. https://doi.org/10.2991/mecae-18.2018.40

[11] Beuter C, Oleskovicz M. S-transform: from main concepts to some power quality applications. IET Signal Process. 2020 May 1;14(3):115-123. https://doi.org/10.1049/iet-spr.2019.0042

[12] Li M, Liu Y, Wang T, Chu F, Peng Z. Adaptive synchronous demodulation transform with application to analyzing multicomponent signals for machinery fault diagnostics. Mech Syst Sig Process. 2023 May 15;191:110208. https://doi.org/10.1016/j.ymssp.2023.110208

[13] Wang X, Ying T, Tian W. Spectrum representation based on STFT. 2020 13th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI); 2020 Oct 17-19; Chengdu (China). Piscataway (NJ): IEEE; 2020. p. 435-438. https://doi.org/10.1109/CISP-BMEI51763.2020.9263516

[14] Alsalmi H, Wang Y. Mask filtering to the Wigner-Ville distribution. Geophysics. 2021 Sep 30;86(6):V489-V496. https://doi.org/10.1190/geo2021-0193.1

[15] Zhu Q, Wang Y, Shen G. Comparison and application of time-frequency analysis methods for nonstationary signal processing. In: Lin S, Huang X, Lin S, Huang XS, editors. Advanced Research on Computer Education, Simulation and Modeling (CESM 2011); 2011; Berlin, Heidelberg. Berlin: Springer Berlin Heidelberg; 2011. p. 286-291. https://doi.org/10.1007/978-3-642-21783-8_47

[16] Bin G, Liao CJ, Li X. The method of fault feature extraction from acoustic emission signals using Wigner-Ville distribution. Adv Mater Res. 2011 Mar 15;216:732-737. https://doi.org/10.4028/www.scientific.net/AMR.216.732

[17] Pan Z, Li C, Zhou X, Deng F, Li T. Combined SPWVD and WVD for high-resolution time-frequency analysis of fault traveling waves. 2025 IEEE 8th International Electrical and Energy Conference (CIEEC); 2025 May 16-18; Changsha (China). Piscataway (NJ): IEEE; 2025. p. 3575-3580. https://doi.org/10.1109/CIEEC64805.2025.11116203

[18] Kalra M, Kumar S, Das B. Target detection using smooth pseudo Wigner-Ville distribution. 2018 IEEE Recent Advances in Intelligent Computational Systems (RAICS); 2018 Dec 6-8; Thiruvananthapuram (India). Piscataway (NJ): IEEE; 2018. p. 6-10. https://doi.org/10.1109/RAICS.2018.8635078

[19] Cao M, Sun Z, Gong E, Xie H. Features extraction of pulse acoustic based on SPWVD. 2017 International Conference on Sensing, Diagnostics, Prognostics, and Control (SDPC); 2017 Aug 16-18; Shanghai. Piscataway (NJ): IEEE; 2017. p. 538-542. https://doi.org/10.1109/SDPC.2017.108

[20] Zhao S, Zhang X, Sun S, Wang X. Research on control law accelerator of digital signal process chip TMS320F28035 for real-time data acquisition and processing. Phys Conf Ser. 2017 Aug 1;887(1):012092. https://doi.org/10.1088/1742-6596/887/1/012092

[21] Bai S, Bai Y. High precision algorithm of metal detector based on balance coil. 2018 21st International Conference on Electrical Machines and Systems (ICEMS); 2018 Oct 7-10; 2018. p. 684-687. https://doi.org/10.23919/ICEMS.2018.8549203