AI FOR SYSTEMATIC REVIEWS, GAIT, AND ACTIVITY RECOGNITION

INTEGRATING AI IN HEALTHCARE: A MULTI-DOMAIN STUDY ON GAIT ANALYSIS, SYSTEMATIC REVIEWS, AND ACTIVITY RECOGNITION

By RONALD LUO, BSc

A Thesis Submitted to the Department of Biomedical Engineering and The School of Graduate Studies of McMaster University in Partial Fulfillment of the Requirements for a Masters of Applied Science Degree.

McMaster University,	Master of Applied Science (2024)
Hamilton, Canada	Biomedical Engineering
TITLE:	Integrating AI In Healthcare: A Multi-domain Study On Gait Analysis, Systematic Reviews, And Activity Recognition
AUTHOR:	Ronald Luo, Honors Biology (Physiology) B.Sc, McMaster University, 2020
SUPERVISOR:	Dr. M. Jamal Deen
NUMBER OF PAGES:	xiii; 113

Lay Abstract

As the global population ages, the importance of addressing healthy aging, healthcare burdens, and continuous vital sign monitoring are becoming increasingly clear. Digital healthcare solutions, such as wearables and smartphone devices, provide a cost-effective, computational means to assess key health indicators.

The initial focus of our research was on gait analysis, specifically investigating whether variations in walking patterns could be effectively used for diagnostic purposes in dementia. Through a systematic review and meta-analysis, we sought to establish a clear link between specific gait patterns and various subtypes of dementia. Our study revealed that distinct gait signatures not only differentiate dementia patients from healthy individuals but also vary significantly across different dementia subtypes, such as Alzheimer's disease, Lewy body disease, frontotemporal dementia, and vascular dementia.

To address the challenges inherent in conducting literature reviews—which are crucial for research teams and entrepreneurs in the medical device and digital technology sectors—we developed software tools to automate the systematic review and meta-analysis process. This development was propelled by the advent of new AI technologies, including the release of OpenAI's models in November 2022. Our software, which utilizes ChatGPT and GPT-3.5 Turbo, automates the inclusion and exclusion of articles, and explores various screening strategies with a focus on improving key performance metrics through an ordinal prompt.

Next, we explored the feasibility of smartphones as health-monitoring tools. Although smartphones contain powerful inertial sensors—accelerometers, gyroscopes, and magnetometers—ensuring consistent data quality and avoiding manual labeling remain significant hurdles. We trained deep learning models to classify human activities from these sensors. While these models achieved strong accuracy in controlled datasets, we observed notable domain shift challenges when applying them to new data, hindering generalization. Nevertheless, our work here underscores the promise of refining data preprocessing and model adaptation techniques to improve consistency and applicability across diverse devices and settings.

Overall, our findings underscore the promise of digital health technologies and AI-driven approaches to streamline research and improve real-world healthcare monitoring. By addressing obstacles in clinical gait analysis, systematic literature reviews, and smartphone-based activity recognition, we lay the groundwork for more accessible and efficient healthcare solutions in an aging society.

Abstract

Background: As the global population ages, the imperative to address challenges related to healthy aging, healthcare burdens, and continuous vital sign monitoring intensifies. Despite numerous digital healthcare solutions, gaps persist in their effective implementation for diagnostic and monitoring applications. This thesis presents innovative tools designed to bridge these gaps, enhancing both the diagnostic capabilities and efficiency of health monitoring technologies.

Objectives: The primary goals of this thesis are threefold: (1) to establish the diagnostic utility of gait signatures for identifying different subtypes of dementia; (2) to enhance the efficiency of systematic literature reviews through automated tools; and (3) to develop and validate models for human activity recognition using smartphone sensor data, addressing challenges such as manual annotation and sensor heterogeneity.

Methods: We conducted a systematic review and meta-analysis to investigate specific gait patterns among individuals with dementia and healthy controls, thereby uncovering subtype-specific gait signatures. We also designed software, leveraging ChatGPT and GPT-3.5 Turbo, to automate critical steps of systematic reviews, optimizing for efficiency and screening performance via different prompting strategies. Finally, we created a smartphone application for gait and activity monitoring, training machine learning models on multi-sensor data to classify daily activities.

Results: Our meta-analysis confirmed that gait parameters reliably distinguish dementia subtypes, including Alzheimer's disease and vascular dementia. The AI-based systematic review tools significantly reduced screening time while maintaining acceptable accuracy, demonstrating the potential of automated evidence synthesis. In the realm of smartphone-based HAR, our models performed robustly in controlled datasets, yet encountered generalization challenges on new data due to sensor heterogeneity and domain shift. These findings highlight both the promise and the practical complexities of scaling smartphone-based monitoring solutions.

Conclusion: Integrating advanced computational approaches and AI into healthcare can enhance diagnostic precision, streamline systematic reviews, and expand smartphone-based health monitoring capabilities. While our models demonstrated strong initial performance, real-world applications require careful consideration of data variability and domain shift. Future research should focus on refining domain adaptation techniques, ensuring more diverse data coverage, and further validating these methods in clinical and community settings.

Acknowledgements

I would first like to express my deepest gratitude to my supervisor, Dr. M. Jamal Deen, for his unwavering support and mentorship throughout my graduate studies. His expertise across biomedical engineering and electrical and computer engineering provided invaluable insights that guided the direction and quality of this research. His willingness to share knowledge and offer encouragement has been crucial to my academic and professional development.

I am also immensely grateful to my committee members, Dr. Qiyin Fang, Dr. Tapas Mandal, and Dr. Sheila Boamah, for their time, constructive feedback, and thesis guidance. Dr. Fang's expertise in biophotonics and Dr. Mandal's medical perspective have both significantly enriched the multidisciplinary nature of this work, helping me refine my methods, interpret data, and articulate the broader impact of the research. I would also like to thank Dr. Boamah for generously stepping in to serve on my defense panel and for her valuable feedback on the final manuscript.

I extend my sincere thanks to my lab colleagues, all of whom are highly skilled medical device researchers in their own right. In particular, I would like to thank Sophini Subramaniam, Abu Ilius Faisal, Ziya Sastimoglu, Mohammad Nami, and Neha Bhattacharyya, whose diverse backgrounds and collaborative spirit propelled our projects forward. Their generosity in sharing expertise, brainstorming solutions, and providing feedback at every stage has been invaluable. I am equally appreciative of my fellow developers and friends, Anson Chang, Saahel Gulistani, and Huiyi Wang, for their dedication to tackling challenging aspects of the mobile sensor application and for keeping morale high during pivotal moments of this journey.

Finally, I would like to thank my parents and my sister for their constant encouragement and belief in me. Their moral and emotional support has been a steady foundation, enabling me to persevere in my work. I could not have completed this thesis without their unwavering confidence in my abilities.

Table of Contents

1. Thesis Introduction 1
1.1 Research Motivations 1
1.2 Extending Healthspan Through Digital Monitoring and Prevention 3
1.3 Evolution of Gait Analysis in Dementia Research 4
1.4 Advancements in Systematic Review Processes 6
1.5 Research Contributions 8
1.6 Thesis Organization 9
2. Systematic Review and Meta Analysis of Gait Signatures in Dementia Subtypes 11
2.1 Introduction 11
2.2 Methods 12
2.3 Results 15
2.4 Discussion 23
2.5 Conclusion 26
3. Large Language Models for Systematic Review Screening and Inclusion 27
3.1 Introduction 27
3.2 Methods 29
3.3 Results 31
3.4 Discussion 36
3.5 Conclusion 39
4. Few-Shot Systematic Review Inclusion with GPT and Scalable Prompt Engineering 40
4.1 Introduction 40
4.2 Methods 42
4.3 Results 44
4.4 Discussion 48
4.5 Conclusion 52

5. Algorithms, Architecture, and Design Patterns for Smartphone-Based Mobility Tracking.....53

5.1 System Overview.....53

5.2 Software Validation Methodologies.....56

5.3 Performance Metrics.....59

5.4 Challenges and Solutions.....62

5.5 Conclusion.....65

6. Deep Learning Models with Smartphone Sensor Data for Human Activity Recognition: Literature Review and Case Study.....66

6.1 Introduction.....66

6.2 Methodology.....67

6.3 Results.....71

6.4 Discussion.....77

6.5 Conclusion.....80

7. Conclusion and Future Directions.....81

7.1 Summary of Key Findings.....81

7.2 Limitations and Challenges.....82

7.3 Future Directions.....83

7.4 Concluding Remarks.....85

List of Figures

FIGURE 2.1 PRISMA FLOW DIAGRAM OF STUDY SELECTION PROCESS 16
FIGURE 2.2 SUMMARY OF STUDY CHARACTERISTICS AND REPORTING TRENDS 17
FIGURE 2.3 EFFECT ESTIMATES OF SPATIOTEMPORAL GAIT PARAMETERS ACROSS DEMENTIA SUBTYPES 18
FIGURE 2.4 NEWCASTLE-OTTAWA SCALE ASSESSMENT OF STUDY RISK OF BIAS 22
FIGURE 2.5 FUNNEL PLOTS ASSESSING PUBLICATION BIAS IN GAIT PARAMETERS 23
FIGURE 3.1 PRISMA DIAGRAM OF GPT-3.5 TURBO SCREENING RESULTS ACROSS REVIEW TOPICS 32
FIGURE 3.2 ALIGNMENT OF MODEL PREDICTIONS WITH HUMAN REVIEWERS 33
FIGURE 3.3 COMPARATIVE ANALYSIS OF GPT-3.5 TURBO MODEL PERFORMANCE ACROSS REVIEWS 34
FIGURE 3.4 ANALYSIS OF FULL-TEXT INCLUSION TASK AND CONTESTED DECISIONS IN AI-ASSISTED REVIEW SCREENING 35
FIGURE 4.1 PRISMA FLOW DIAGRAM AND CONFUSION MATRICES SET AT OPTIMAL THRESHOLD FOR EACH REVIEW 45
FIGURE 4.2 ROC CURVES FOR EVALUATING OPTIMAL ORDINAL THRESHOLD COMPARED TO RANDOM MULTI-CLASS CONTROL 46
FIGURE 4.3 CONFUSION MATRICES SET AT AVERAGE THRESHOLD AND HISTOGRAM OF CONFIDENCE SCORES 47
FIGURE 5.1 ARCHITECTURE COMPONENT DIAGRAM 54
FIGURE 5.2 REGISTRATION, VERIFICATION, AND AUTHENTICATION FLOW 54
FIGURE 5.3 CLIENT BUFFERING AND STORAGE INFRASTRUCTURE 55
FIGURE 5.4 ENTITY RELATIONSHIP DIAGRAM OF POSTGRES DATABASE 57
FIGURE 5.5 SENSOR RESPONSES DURING FREE FALL AND IMPACT ACROSS TEST DEVICES 61
FIGURE 5.6 SENSOR DATA DURING 10-METER WALKING TRIALS WITH SMARTPHONE IN POCKET 63
FIGURE 5.7 APPLICATION SCREENS 64
FIGURE 5.8 USER INFO AND CONSENT FLOW 65
FIGURE 6.1 DEEP LEARNING MODEL ARCHITECTURES 69
FIGURE 6.2 YEAR OF PUBLICATION PLOTTED AGAINST ARTICLE TOPIC RELEVANCE 72
FIGURE 6.3 RESULTANT SMARTPHONE SENSOR DATA SORTED BY ACTIVITY 74
FIGURE 6.4 COMPARISON OF ACCELEROMETER BASED PERFORMANCE USING CONFUSION MATRICES AND MODEL METRICS 75
FIGURE 6.5 COMPARISON OF GYROSCOPE BASED PERFORMANCE USING CONFUSION MATRICES AND MODEL METRICS 76
FIGURE 6.6 TRAINING AND SUPPLEMENTARY DISTRIBUTIONS OF ACCELEROMETER AND GYROSCOPE SENSOR READINGS 77
FIGURE 6.7 TEST SET RESULTANT SENSOR READINGS AND CONFUSION MATRICES 78

List of Tables

TABLE 1.1 ADVANCEMENTS IN GAIT ANALYSIS DEVICES OVER THE PAST TWO DECADES. 5
TABLE 1.2 GAIT ANALYSIS OUTCOMES OVER THE PAST TWO DECADES. 6
TABLE 1.3 TYPES OF SYSTEMATIC REVIEWS IN BIOMEDICAL RESEARCH 7
TABLE 2.1 KEYWORDS USED IN DATABASE SEARCH 13
TABLE 2.2 SUMMARY EFFECT SIZES OF SINGLE-STUDY DEMENTIA SUBTYPES SPATIOTEMPORAL PARAMETERS 19
TABLE 2.3 SUMMARY EFFECT SIZES OF DEMENTIA SUBTYPES SPATIOTEMPORAL PARAMETERS FROM META-ANALYSIS 19
TABLE 3.1 COMPARATIVE ANALYSIS OF GPT-3.5 TURBO'S PERFORMANCE METRICS FOR TITLE-ABSTRACT INCLUSION ACROSS REVIEW TOPICS 33
TABLE 3.2 COMPARATIVE ANALYSIS OF GPT-3.5 TURBO'S PERFORMANCE METRICS FOR FULL-TEXT INCLUSION ACROSS REVIEW TOPICS 36
TABLE 4.1 BINARY AND ORDINAL PROMPTS USED TO ELICIT MODEL RESPONSES 43
TABLE 4.2 PERFORMANCE METRICS COMPARING OPTIMAL ORDINAL THRESHOLD TO BINARY PROMPT 46
TABLE 4.3 PERFORMANCE METRICS OF AVERAGE THRESHOLD FOR ORDINAL COMPARED TO BINARY PROMPT 48
TABLE 4.4 RATIONALE FOR INCLUSION SCORES ASSIGNED BY GPT-3.5 TURBO IN SYSTEMATIC REVIEW SCREENING OF LONG-COVID HEALTH OUTCOMES 49
TABLE 5.1 SAMPLING RATE DISCREPANCIES FOR SMARTPHONE IMU SENSORS AT DIFFERENT INPUT FREQUENCIES 60
TABLE 6.1 KEYWORDS USED IN DATABASE SEARCH 68

List of Abbreviations

2FA 2-Factor Authentication
AD Alzheimer's Disease
AI Artificial Intelligence
ANN Artificial Neural Network
ANOVA Analysis of Variance
API Application Programming Interface
AUC Area Under Curve
CDC Centers for Disease Control and Prevention
CI Confidence Interval
CIHR Canadian Institute for Health Research
CNN Convolutional Neural Network
COVID Coronavirus Disease
CTA Call To Action
CV Coefficient of Variation
DNN Dense Neural Network
ECG Electrocardiogram
EMG Electromyography
FPR False Positive Rate
FFN Feed Forward Network
FTD Frontotemporal Dementia
GDP Gross Domestic Product
GPS Global Positioning System
GPT Generative Pre-trained Transformer
GRU Gated Recurrent Unit
HAR Human Activity Recognition
HNSW Hierarchical Navigable Small World
HTTP Hypertext Transfer Protocol
IMU Inertial Measurement Unit
JBI Joanna Briggs Institute System
LBD Lewy Body Dementia
LDA Latent Dirichlet Allocation
LLM Large Language Model

LSTM Long Short Term Memory
MCC Mathew's Correlation Coefficient
MeSH Medical Subject Heading
ML Machine Learning
MOOSE Meta-Analyses and Systematic Reviews of Observational Studies
NIH National Institute of Health
NLP Natural Language Processing
NOS Newcastle-Ottawa Scale
NPV Negative Predictive Value
PK Primary Key
PPV Positive Predictive Value
PRISMA Preferred Reporting Items for Systematic reviews and Meta-Analyses
RAG Retrieval Augmented Generation
ReLU Rectified Linear Unit
RF Random Forests
RNN Recurrent Neural Network
ROC Receiver Operating Characteristic
SMD Standard Mean Difference
SMS Short Message Service
SpO2 Arterial Oxygen Saturation
SQL Structured Query Language
SVM Support Vector Machine
TF-IDF Term Frequency-Inverse Document Frequency
TPR True Positive Rate
TUG Timed Up and Go Test
UCI University California Irvine
UI/UX User Interface and Experience
VaD Vascular Dementia
WHO World Health Organization

List of Symbols

P	Significance Value
d	Cohen's Effect Estimate
X^2	Chi-Squared Statistic
I^2	Heterogeneity Statistic
k	Number of Included Studies
n	Number of Articles (Chapter 2, Chapter 3)
n_D	Number of Dementia Patients (Chapter 2)
n_C	Number of Healthy Controls (Chapter 2)
N	Number of Articles (Chapter 4)
T	Threshold of Confidence Score for Article Inclusion
W	Workload Reduction
S	Inclusion Suitability Score
f	Sampling Frequency
Δs	Sampling Interval
n	Total Number of Samples (Chapter 5)
t	Total Elapsed Time of Recording
a	Estimated Linear Acceleration Value
g	Gravitational Acceleration Constant
s	Sensor Scaling Factor Reading
R	Resultant Vector Magnitude
x	X-Axis Reading
y	Y-Axis Reading
z	Z-Axis Reading
Z_i	Indexed Z-Score Normalization Value
x_I	Indexed Sensor Reading Value
$\bar{x}$	Mean Sample Sensor Value
σ	Mean Standard Deviation Value

1. Thesis Introduction

In the late 20th century, concerns about overpopulation were prominently featured in discussions and literature, notably in "The Population Bomb," a 1968 book by Stanford University professor Paul R. Ehrlich. This publication echoed Malthusian theory, which posited that population growth would outpace agricultural production, leading to widespread scarcity and hardship. While Malthusian predictions influenced policies like China's one-child policy, aimed at controlling population growth, these measures were part of a broader global response to the perceived consequences of unchecked population expansion.

Today, however, the global demographic landscape tells a different story. Instead of overpopulation, many regions face what could be termed an "underpopulation cliff," characterized by declining birth rates and increased life expectancy. This shift is particularly pronounced in countries like Japan, where the number of older adults is projected to exceed that of young adults by as early as 2030. This trend threatens the balance of social support systems, as a smaller working-age population is left to support a growing number of elderly.

In this chapter, we outline the motivations for this thesis, beginning with the pressing demographic and economic challenges associated with aging populations. We introduce the disparities between lifespan and healthspan, and explore the ongoing advancements in gait monitoring and digital wearable technologies. Next, we describe the current state of the art in implementing health policy and practice through systematic reviews, and distinguish these methods from more recent developments. Lastly, we highlight the contributions of our research and the structure of this thesis, indicating how our work intersects with these critical areas and serves to address key challenges.

1.1 Research Motivations

1.1.1 Global Responses to Aging Populations, Policy Initiatives and Research Funding: In developed countries, increased life spans have been largely achieved through advancements in medical science, enhanced childhood nutrition, and prolonged periods of peace [1], [2], [3]. These gains reflect significant progress in policy, public health, and technological development. However, these extended lifespans have been accompanied by declining birth rates, shifting demographic structures dramatically. Such demographic changes pose substantial challenges, as younger populations are crucial for sustaining labor markets, infrastructure development, pension contributions, and overall economic health [4], [5].

The trend towards an aging population carries significant public health implications as well, evident in the evolving demographic pyramids of various countries. In nations with large aging populations like Japan, Italy, and Germany, the pressure on healthcare systems is becoming increasingly

apparent [6], [7], [8], [9]. These pressures are characterized by rising healthcare costs, greater proportions of Gross Domestic Product (GDP) being allocated to healthcare services, and growing national debt driven by the need to support expansive social welfare systems [10].

Governments and international organizations are actively responding to these demographic shifts with a range of policy initiatives and increased funding for aging-related research. Japan's "Society 5.0" initiative [11], for example, integrates advanced technologies like robotics and Artificial Intelligence (AI) to enhance elderly care. Similarly, countries like Germany and Italy are overhauling their pension systems and investing in community-based services to support active aging and manage healthcare expenditures effectively [12].

On a broader scale, the European Union, through its Horizon Europe program, is increasing support for research aimed at improving elderly life quality and the sustainability of health systems [13]. In North America, the U.S. National Institute on Aging (NIA) and Canada's Canadian Institutes of Health Research (CIHR) have expanded their funding focuses to encompass geriatrics and longevity studies, with CIHR's "Healthy Aging" initiative specifically aimed to enhance the health span of older adults and support age-friendly communities [14], [15]. In Canada, seniors aged 65 and older represent nearly one in five people, a proportion expected to rise to nearly a quarter by 2030 [16]. This growing demographic is already placing pressure on healthcare and long-term care systems, prompting nationwide investments in assistive technologies, telehealth, and age-inclusive policy reforms. In response, governments and research agencies are increasingly exploring digital health tools, such as wearable sensors, remote monitoring systems, and artificial intelligence, as part of their aging strategies. These innovations offer not only a means to deliver more scalable care but also an opportunity to accelerate the translation of research into real-world clinical practice.

1.1.2 Accelerating Health Research Translation with AI and Digital Technologies: Despite the growing number of global initiatives aimed at addressing the challenges posed by aging populations, a critical gap remains in the speed at which research findings are translated into clinical practice and policy. Literature in implementation science highlights that the conventional pathway from discovery to practice extends up to 17 years, which is at increasing odds with the urgent needs of today's rapidly aging societies [16], [17]. This research seeks to expedite this process by leveraging advancements in digital wearables and science implementation technology. It focuses on developing practical solutions for gait monitoring, activity recognition, and automating the systematic review process—each a critical piece in addressing current demographic shifts.

1.2 Extending Healthspan Through Digital Monitoring and Prevention

1.2.1 Introduction to Lifespan vs. Healthspan: The need to differentiate between 'lifespan'—the total number of years one lives—and 'healthspan'—the period during which an individual remains healthy and active—is crucial in addressing the challenges of an aging population [18], [19], [20]. While medical advancements have successfully extended lifespan, they have not equally improved healthspan. Many older adults endure debilitating conditions that impair their mobility, diminish their sense of agency, and reduce their overall life satisfaction. These issues are often compounded by chronic comorbidities, making the latter years of life less fulfilling and significantly more costly in terms of healthcare [21], [22].

1.2.2 Economic Implications of the Healthspan Gap: From an economic perspective, the costs associated with long-term care, chronic disease management, and end-of-life care are immense and growing, reflecting a significant burden on healthcare systems worldwide. For instance, the global cost of dementia care, including Alzheimer's disease (AD), Lewy body dementia (LBD), frontotemporal dementia (FTD), and vascular dementia (VaD), is projected to reach 1.7 trillion by 2030, driven by the needs for continuous medical treatment, support systems for patients and caregivers, and specialized residential care facilities [23], [24]. These escalating costs highlight the urgent need for effective management strategies that extend healthspan and reduce the financial strain on public and private healthcare sectors.

The World Health Organization (WHO) estimates that every dollar spent scaling up interventions to prevent noncommunicable diseases returns multiple folds in health cost savings and productivity gains [25]. Effective public health programs, such as childhood vaccinations and tobacco control, have demonstrated substantial economic benefits. For instance, the global polio eradication initiative has significantly reduced healthcare costs worldwide [26]. In the United States, routine childhood immunizations are estimated by the Centers for Disease Control and Prevention (CDC) to save about $3.24 trillion in direct and societal costs for cohorts born between 1994 and 2023 [27]. Additionally, California's tobacco control program has dramatically lowered hospital admissions for heart attacks and strokes [28], saving billions in healthcare costs. These examples highlight the profound impact of preventive healthcare on both health outcomes and economic burdens. As these conditions progress, they often lead to notable declines in physical function, particularly mobility and balance, making them prime candidates for monitoring through digital tools like gait analysis.

1.2.3 Technological Innovations as Solutions: Just as policy adjustments are critical in managing the economic challenges of aging populations, technological innovations in healthcare represent a parallel strategy that has proven effective in mitigating health and financial burdens [29]. Research by [30] in Nature Reviews Cardiology demonstrates how smart wearables can reduce hospital readmissions for patients with coronary artery disease by facilitating continuous monitoring and timely interventions. Likewise, a study by [31] shows that wearables and digital devices can enhance medication adherence through timely reminders, thus minimizing future healthcare expenditure.

Finally, the proliferation of computing devices has significantly enhanced the scope of digital health monitoring. State-of-the-art AI models, such as Convolutional Neural Networks (CNNs), now rival dermatologists in detecting melanoma—a capability enhanced by the widespread availability of smartphone cameras [32], [33]. Similarly, SpO2 sensors incorporated in smartwatches provide vital data on blood oxygen levels, aiding in the detection of respiratory or cardiovascular anomalies [34], [35]. These advancements and more exemplify a broader trend toward proactive and preventive healthcare, leveraging everyday technology to improve health outcomes and reduce systemic costs.

1.3 Evolution of Gait Analysis in Dementia Research

1.3.1 Connection Between Gait and Neurological Health: Gait serves as a functional indicator of neurological health due to its reliance on the brain's complex coordination. Walking engages multiple brain regions, including the motor cortex, cerebellum, and basal ganglia, each critical for orchestrating fluid motor activities [36]. Disturbances in these areas due to neurological disorders manifest as distinct changes in gait. For example, cerebellar neurodegeneration can lead to ataxia, characterized by unsteady and uncoordinated movements [37]. Similarly, impairments in the frontal lobe, often associated with various types of dementia, can affect executive functions and motor control, resulting in altered stride length, speed, and symmetry. Recent studies have even elevated the importance of walking speed, proposing it as the 'sixth vital sign' due to its strong correlation with critical health outcomes such as survival rates and recovery success [38].

1.3.2 Historical Overview and Technological Advancements*:* The well-established connection between neurology and gait analysis underpins our laboratory's research focus [39], [40], [41], [42]. Gait analysis has been a fundamental tool in neurology for decades, offering critical insights into the functionality of the central nervous system [43]. Traditionally, gait irregularities associated with neurological disorders like Parkinson’s disease, multiple sclerosis, and stroke were observed visually by clinicians [44]. However, methodological advancements have led to the adoption of more sophisticated techniques such as electronic walkways [45] and motion capture systems [46].

These improvements, driven by advancements in computational hardware and software, have significantly enhanced the precision of gait analysis systems [47], [48], [49], [50]. Modern gait analysis devices now employ a diverse array of sensors, wearables, and software algorithms to capture and analyze detailed gait parameters. Such technological progress has facilitated the detection of minute gait abnormalities, previously imperceptible, thereby enabling earlier and more accurate predictions.

1.3.3 Heterogeneity in Gait Analysis Instrumentation and Outcome Data: The evolution of gait analysis tools over the past two decades is marked by increasingly sophisticated devices that offer greater precision and integration—from simple pedometers to complex motion capture systems that provide 3D kinematic and kinetic feedback. A table summarizing key advancements in gait analysis devices illustrates how each generation of technology has enhanced the utility and effectiveness of gait analysis in healthcare settings.

Table 1.1 Advancements in Gait Analysis Devices Over the Past Two Decades.

Device Type	**Technological Features**	**Application in healthcare**
Pedometers	Basic Step Counting	Initial quantitative assessment of walking activity
Electronic Walkways	Pressure sensors, integrated timing	Detailed gait cycle analysis
Wearable Sensors	Accelerometers, gyroscopes	Real-time gait monitoring
Motion Capture Systems	3D kinematic and kinetic analysis	Comprehensive gait analysis in research labs
Smart Insoles	Pressure distribution mapping, wireless data transmission	Continuous monitoring, feedback for gait correction

In addition to the diversity in instrumentation, gait analysis also encompasses a wide range of research focuses and outcomes that vary among studies. Clinicians and researchers often measure various aspects of mobility through different protocols, such as activity recognition, spatiotemporal parameters, the timed up and go test, as well as tasks performed under single and dual-task conditions. The table below demonstrates the variety of different gait analysis methods.

Table 1.2 Gait Analysis Outcomes Over the Past Two Decades.

Outcome Type	Measurement Protocol	Relevance to Clinical Research
Spatiotemporal Parameters	Stride length, speed, cadence	Fundamentals of gait mechanics
Timed Up and Go (TUG) Test	Time taken to stand up, walk, return, sit down	Mobility and balance assessment
Single-task Gait Testing	Walking while performing no additional tasks	Baseline mobility assessment
Dual-task Gait Testing	Walking while performing a cognitive or motor task	Cognitive-motor interference assessment
Activity Recognition	Machine learning models analyzing sensor data	Classifying types of physical activity

1.3.4 From Laboratory to Everyday Use: Despite the advancements and diversification in gait analysis tools, traditional gait analysis has often been overlooked and remains confined to specialized settings with costly equipment. A shift to using ubiquitous technologies like smartphones, which are equipped with sensors such as accelerometers and gyroscopes, would make gait analysis more accessible, facilitating its application in more naturalistic and everyday environments. In later chapters, we will discuss the development of a smartphone application for gait and activity monitoring, an advancement that utilizes everyday technology to support health monitoring. This transition expands the potential for continuous, real-life data collection, which would enable early diagnosis and real-time disease progression monitoring. We also tackle challenges like manual data annotation and sensor heterogeneity, showing how AI can simplify and improve the efficiency of health data analysis.

1.4 Advancements in Systematic Review Processes

1.4.1 Introduction to Systematic Reviews in Biomedical Research: As we explore advancements in wearables and digital health technologies, it becomes essential to address how these innovations are integrated into healthcare practice. Systematic reviews, positioned at the top of the evidence hierarchy, are crucial for health research translation. These comprehensive analyses synthesize vast quantities of research publications, establishing clinical guidelines, informing policy decisions, and identifying areas needing further investigation [51]. Given the exponential growth in biomedical research, systematic reviews are more critical than ever in managing the deluge of information and ensuring that findings are accurately incorporated into medical practice and policy development.

1.4.2 Variability in Systematic Review Approaches: Systematic reviews vary significantly in scope, methodology, and application, allowing researchers and policymakers to select a review type that best fits

their specific needs. For instance, a meta-analysis may be used to derive precise effect sizes from quantitative data in clinical trials, while a scoping review could be used to map out the breadth of research and identify gaps in a newly emerging field [52], [53], [54]. To illustrate the range and utility of these review types, the following table categorizes some of the most common systematic reviews utilized in biomedical research, showcasing their distinct methodologies and applications.

Table 1.3 Types of Systematic Reviews in Biomedical Research

Review Type	**Description**
Narrative Reviews	Provide a qualitative synthesis and broad overview of a field, using less structured methodologies.
Scoping Reviews	Aim to map the key concepts, types of evidence, and main sources in a research area, often to identify gaps.
Meta-Analyses	Combine results from quantitative studies statistically to offer a precise estimate of treatment effects or risks.
Rapid Reviews	Offer a quicker, less comprehensive analysis, often employed in policy making during urgent situations.

1.4.3 Evolution of Systematic Review Methodologies and Key Contributors: Systematic review methodologies are entrenched in the principles of evidence-based medicine, with the Cochrane Collaboration playing a pivotal role since its establishment in 1993. To date, the Cochrane Handbook has provided the most comprehensive and up-to-date guidelines on conducting extensive database searches, managing study selection, extracting data, and analyzing results [55]. Building on the Cochrane protocol, the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines have improved the transparency and reproducibility of systematic reviews by mandating detailed flow diagrams for documentation–now a standard requirement for many journal submissions [56], [57]. For the specific needs of nursing and allied health fields, the Joanna Briggs Institute (JBI) tailors its protocols to encompass various forms of evidence synthesis, including guidelines for scoping reviews [58]. Meanwhile, the Meta-analysis Of Observational Studies in Epidemiology (MOOSE) protocol focuses on observational studies and prescribes specific criteria for study selection and validity assessment that maintains the integrity of the overall findings [59].

Technological advancements have complemented these structured methodologies to further streamline the systematic review process. At the outset, PROSPERO allows researchers to pre-register their review protocols, promoting transparency and reducing redundancy by preventing duplication. In the intermediary stage, tools like Covidence and Rayyan help with study screening and data management,

allowing multiple researchers to work on a single dataset [60]. Towards the final stages, Review Manager (RevMan, Cochrane Collaboration) handles data analysis and visualization [61]. The software supports the meta-analysis phase by providing statistical tools for calculating effect sizes, such as standard mean difference, odds ratio and risk ratio.

1.4.4 Systematic Review Challenges and the Transformative Impact of AI Technologies: Despite the advancements in systematic review methodologies and supporting technologies, significant challenges persist. Traditional systematic reviews involve extensive manual efforts, from literature search and screening to data extraction and quality assessment [62]. Each step must be meticulously performed to ensure the reliability of the review, often requiring multiple reviewers and months of work [63]. This intensive nature poses significant barriers, particularly when rapid syntheses of research findings are necessary, as seen during health emergencies like pandemics [54]. Moreover, despite efforts to mitigate bias, the subjectivity inherent to manually screening thousands of studies can lead to inconsistencies in study selection and data interpretation, potentially affecting the outcome of the review.

Given the extensive challenges inherent in traditional systematic review processes, the integration of AI, machine learning (ML) and natural language processing (NLP) offers a transformative shift. Initially, researchers tried techniques such as support vector machines (SVMs) and random forests (RFs) to automate parts of the screening process [47], but these early models often struggled with accuracy and scalability [64], [65]. Recently, it has become feasible to employ more sophisticated methods, driven by the release of Large Language Models (LLMs) by organizations such as OpenAI and Hugging Face [66], [67]. Looking ahead, as computational power becomes more accessible and model capabilities expand, extending AI to tasks such as literature screening, data extraction, and even the qualitative synthesis of studies will significantly reduce the time and labor required for research translation.

1.5 Research Contributions

1.5.1 A Convergent Approach to Digital Health: This thesis explores interconnected projects that advance healthcare delivery and research through the application of emerging technologies. Focused on developing practical solutions, this research aims to bridge the gap between lifespan and healthspan, aligning technological advancements with the immediate needs of an aging global population. The key contributions below highlight the impact and scope of this work.

1.5.2 Systematic Review and Meta-Analysis of Gait Signatures in Dementia Subtypes: This project uniquely contributes through the systematic review and meta-analysis of gait parameters across different dementia subtypes. Our findings revealed distinct gait patterns that could improve diagnostic precision

and facilitate the differentiation of subtypes. Notably, while the majority of recent studies have centered on Alzheimer's Disease (AD), our research highlighted a substantial gap concerning non-AD dementia subtypes. These subtypes remain understudied and susceptible to misdiagnosis, underscoring the urgent need for more focused research and increased funding to enhance diagnostic capabilities.

1.5.3 Development and Validation of LLM Tools for Systematic Review Automation: This work uniquely contributes through the development and testing of automation tools for systematic review using low-cost LLMs. Our web application used Retrieval Augmented Generation (RAG) to automate several stages of article screening and data extraction. We evaluated the performance of the tool by comparing GPT-3.5 Turbo's responses with manual responses obtained by human reviewers across three different systematic review topics. Notably, the model exhibited considerable efficiency gains and showed promise for handling both title-abstract and full-text articles, achieving high performance across a broad spectrum of evaluation metrics. Further work demonstrated the capability of specialized prompts to maximize the sensitivity and specificity of the generated responses.

1.5.4 Smartphone Application for Gait and Activity Monitoring: This final project marks a significant contribution through the development of a smartphone application that captures naturalistic human gait data. Our application uses IMUs embedded in smartphones to collect data on user movement patterns and activities. We benchmarked our system against established market IMUs, and incorporated machine learning models—including densely connected networks, Long Short-Term Memory (LSTM), and Gated Recurrent Units (GRU)—to analyze data from embedded IMUs. Our proposed system demonstrated remarkable distinguishing power on training data, achieving high performance across multiple evaluation metrics, and shows strong promise for real-world, gait-based health monitoring applications.

1.6 Thesis Organization

1.6.1 Chapter 1. Introduction and Research Motivations: This chapter sets the stage for the thesis by introducing the evolving demographic trends, particularly the challenges posed by aging populations and the resultant health and economic impacts. It details the motivation behind the research, highlighting the need for advancements in health technology to address the disparities between lifespan and healthspan. The chapter also contextualizes the role of systematic reviews in translating health research into practice, paving the way for a discussion on the integration of AI technologies.

1.6.2 Chapter 2. Dementia and Gait Systematic Review and Meta-Analysis: This chapter presents the systematic review and meta-analysis that underpins the thesis, outlining the methodology used to select and analyze studies investigating gait patterns in dementia patients versus healthy controls. The findings

from this review show signs of distinctive gait signatures associated with various dementia subtypes, highlighting the diagnostic potential of gait analysis.

1.6.3 Chapter 3. Evaluating the Efficacy of Large Language Models for Systematic Review and Meta-Analysis: Chapter 3 investigates the development and validation of AI tools aimed at automating the systematic review process. This section examines the use of LLMs to improve the efficiency and accuracy of literature screening, presenting comparative studies of AI and human performance in screening tasks.

1.6.4 Chapter 4. Advancing Few-Shot Systematic Review Inclusion with GPT and Scalable Prompt Engineering: Building on the previous chapter, Chapter 4 examines advanced techniques in AI, specifically focusing on few-shot learning and scalable prompt engineering within GPT models. This chapter showcases how these technologies can further refine the systematic review process, making it even more efficient and adaptable to various scenarios.

1.6.5 Chapter 5. Smartphone Application for Human Activity Recognition: Chapter 5 describes the development and implementation of the smartphone application designed to monitor gait and other physical activities. It covers the technical aspects of the app, including the use of smartphone sensors and the algorithmic processing of sensor data. The chapter also discusses the results of preliminary trials and the app's potential in real-world settings for early diagnosis and continuous monitoring of neurological health.

1.6.6 Chapter 6. Smartphone Sensor Data for Human Activity Recognition: Building on Chapter 5, Chapter 6 conducts a literature review of automated labeling methods and evaluates three deep learning models—feed-forward neural networks, LSTM, and GRU—using publicly available datasets. While these models show strong performance in controlled settings, the chapter also highlights the challenges of domain shift and sensor heterogeneity when applying them to new datasets.

1.6.7 Chapter 7. Conclusion and Future Directions: The final chapter synthesizes the findings from all previous chapters, discussing the overall contributions of this thesis to the field of digital health and biomedical engineering. It reflects on the limitations encountered during the research and proposes future directions for expanding the scope of AI applications in healthcare, particularly in managing and monitoring age-related conditions.

2. Systematic Review and Meta Analysis of Gait Signatures in Dementia Subtypes

As established in the Introduction, this thesis aims to bridge multiple domains—gait analysis, systematic reviews, and human activity recognition (HAR)—through the lens of modern computational tools and AI-driven approaches.

The work in Chapter 2 serves as a critical baseline: it not only illustrates the complexity of extracting meaningful signals from clinical studies but also lays the groundwork for subsequent chapters, which will tackle the methodological hurdles of evidence synthesis and activity monitoring. Building on the broader motivations outlined in Chapter 1, this chapter focuses on identifying objective, quantifiable gait metrics that could differentiate dementia subtypes—such as Alzheimer's disease, Lewy body dementia, frontotemporal dementia, and vascular dementia. By systematically reviewing and synthesizing the existing literature, we attempt to create a foundation of evidence that will later inform more technologically driven solutions.

Having identified where the challenges lie—such as heterogeneity in measurement, differences in instrumentation, and the need for robust data analysis methods—this chapter paves the way for the innovations in systematic review automation and wearable solutions that follow in Chapters 3 through 6. It sets an anchor point that aligns clinical questions with practical patient outcomes [*].

2.1 Introduction

In 2020, over six million Americans of all ages were living with Alzheimer's disease (AD), according to the Alzheimer's Association [68]. Globally, an estimated 55 million people have dementia, and projections indicate that the number of dementia cases in the U.S. could nearly double, rising from 514,000 in 2020 to nearly 1 million by 2060 [69].

Dementia manifests in various forms, each with distinct symptoms and underlying causes. These symptoms can include memory loss, mood changes, communication difficulties, impaired executive function, and behavioral changes. Among the wide range of symptoms, movement disorders — particularly those associated with extrapyramidal signs — are notable [70]. Gait disturbances are also common, but can be caused by factors like vascular pathology, physical injury, or sensory deficits, often complicating the diagnosis by masking primary neurological symptoms. As a result, diagnosing dementia

[*] In consideration for IEEE Journal of Biomedical and Health Informatics (Impact Factor 6.7)

remains a complex process, requiring a combination of medical history, physical examination, laboratory tests, and neuroimaging, as there is no single definitive test for dementia [71].

Vascular dementia (VaD) and Lewy body dementia (LBD), each account for about 20% of dementia cases, while Alzheimer's disease (AD) constitutes approximately 60% of all cases [72]. Frontotemporal dementia (FTD) is rarer, with an incidence of roughly 3.4 per 100,000 people [73]. Together, these four subtypes make up the majority of dementia cases. However, due to overlapping symptoms, misdiagnosis is common, with non-AD dementias sometimes mistaken for AD. Accurate early diagnosis is crucial, as each subtype may respond differently to treatments [72], [74].

Recent studies have highlighted gait deficits in the early stages of dementia, suggesting that gait analysis could be a useful diagnostic tool [70]. Advances in measuring techniques have identified distinct gait patterns across dementia subtypes. Recognizing these gait signatures offers the potential for a non-invasive, easily administered diagnostic tool that could not only aid in early diagnosis and disease progression tracking, but also assist with financial planning, living arrangements, and legal matters for individuals with dementia [75].

While previous research has explored gait differences among dementia subtypes, there remains a lack of quantitative meta-analyses, particularly regarding LBD [76], [77]. This study aims to compare spatiotemporal gait parameters (which describe both spatial aspects, such as stride length, and temporal aspects, such as cadence) across the four most prevalent dementia subtypes (AD, LBD, FTD, and VaD), using data from instrumented gait analysis tools. Our goal is to determine whether distinct gait patterns can differentiate between dementia subtypes and healthy controls, addressing the need for more precise diagnostic tools in dementia care.

Building on our laboratory's previous work, especially in wearable health analyzers for gait analysis [39], [40], [41], [42], [78] this systematic review and meta-analysis seeks to expand the knowledge base surrounding dementia diagnosis. By evaluating various instrumental approaches, our study could contribute to the development of more effective diagnostic tools for dementia, ultimately improving patient outcomes and quality of life.

2.2 Methods

2.2.1 Study Design: This systematic review and meta-analysis adhered to the study design guidelines recommended by the *Cochrane Collaboration* [55] and followed the *Guidelines for Meta-Analysis and Systematic Reviews of Observational Studies* (MOOSE) [59]. The study protocol was pre-registered with the *International Prospective Register of Systematic Reviews* (PROSPERO, Registration Number:

CRD42022295839). Additionally, the study followed the *Preferred Reporting Items for Systematic Reviews and Meta-Analysis* (PRISMA) statement [57]. As the study only used published data and did not involve confidential participant information or interventions, no ethical approval was required. The search timeline was adjusted from the original protocol due to logistical constraints, with an updated search being conducted in November 2024 instead of January 2022.

2.2.2 Search Strategy: A comprehensive literature search was conducted in November 2024 across multiple databases, including Embase, CENTRAL, PubMed, Web of Science, and CINAHL. These five databases were chosen for their strong coverage of biomedical and dementia-related research. The search strategy was developed with input from a health sciences librarian. The search covered publications from 1974 to 2024. A combination of relevant keywords and MeSH terms related to dementia and gait analysis were used, and the full search strategy can be found in Table 2.1 below. Title and abstract screening were carried out independently by two reviewers (R.L. and S.S.), with discrepancies resolved independently by a third reviewer (A.I.F.). Full-text screening followed the same process, with different pairs of reviewers (R.L. and A.I.F.). The research was overseen by experts in biomedical engineering and health research methods (L.M. and M.J.D.).

Table 2.1 *Keywords Used in Database Search*

Category	Keywords
Neurological	"Alzheimer disease", "Alzheimer*", "dementia", "frontotemporal dementia", "lewy body", "diffuse Lewy body disease", "vascular dementia", "neurologic disease"
Motor	"gait analysis", "gait disorder", "mobility", "walk*", "walking"
Spatiotemporal Parameters	"cadence", "double support", "walking distance", "gait characteristics", "gait parameters", "range of motion", "stance time", "stride length", "stride velocity", "step time", "step length", "swing time", "single support", "single task", "walking speed"

Note: The database search used OR operators within each concept and AND operators between concepts to refine results. MeSH (Medical Subject Headings) were used where possible, while keywords were applied across ALL FIELDS to maximize relevant study retrieval.

2.2.3 Selection Criteria: The primary outcomes of interest were spatiotemporal gait parameters derived from natural walking and instrumented analysis. Studies eligible for meta-analysis were required to report descriptive statistics, such as mean, standard deviation, and sample size. All study designs were considered, provided the participants included individuals with AD, FTD, LBD, or VaD. Studies involving comorbid conditions were included, as per the protocol. For experimental studies, only baseline

measurements were analyzed. Only peer-reviewed English-language publications were included, while books, conference abstracts, and animal studies were excluded. Studies involving only treadmill walking or dual-task assessments were excluded based on protocol guidelines. Systematic reviews identified during screening were used to perform forward reference searching for additional relevant studies.

2.2.4 Data Extraction: Study details (e.g., author, year, design), instrument used, walking protocol, dementia subtype, sample size, and key spatiotemporal gait parameters (e.g., cadence, gait velocity, stride length) were similarly extracted in pairs using a standardized form. These spatiotemporal parameters were standardized to common units for comparison. The protocol initially aimed to retrieve missing data by contacting study authors. In practice, sufficient data were available from extracted studies, and contacting authors for missing data was not required. Missing values were estimated from available statistics or excluded entirely.

2.2.5 Data Synthesis, Analysis, and Statistics: Meta-analysis was performed to compare gait parameters between dementia subtypes and healthy controls, calculating standardized mean differences (SMD). Statistical analysis was conducted using *Review Manager* (RevMan 5.4) [61] with a significance level set at $P < 0.05$. Results were presented as forest plots with 95% confidence intervals, and effect sizes were interpreted using *Cohen's* criteria (small: $d > 0.20$; medium: $d > 0.50$; large: $d > 0.80$) [79].

Subgroup analyses were performed based on dementia subtypes and geographical locations. Age-, sex-, and severity-based subgroup analyses were not conducted due to inconsistencies in the reported data across studies. Heterogeneity was assessed using the *Cochrane* X^2 test and the I^2 *statistic*. A random-effects model was applied due to anticipated variability across studies. Potential sources of heterogeneity, including measurement methods, geographic regions, and risk of bias (assessed using the *Newcastle-Ottawa Scale*), were considered. Reporting bias was assessed using funnel plots and Egger's tests.

2.2.6 Study Risk of Bias Assessment: The *Risk of Bias* (ROB) of the included studies was assessed using the *Newcastle-Ottawa Scale* (NOS) for case-control studies [80], which evaluates selection, comparability, and exposure criteria, with a maximum score of eight stars. Although the protocol initially planned for the use of NOS for cohort studies, the case-control version was ultimately deemed most appropriate based on the study designs included in this review. The key domains assessed included case representativeness, control selection, exposure ascertainment, and comparability of groups. Two authors (R.L. and S.S.) independently conducted the *ROB* assessments, with a third reviewer (A.I.F.) independently resolving discrepancies. The results were presented in a stacked bar chart format.

Sensitivity analyses, initially planned to evaluate both study design and *ROB*, were limited to assessing bias risk due to data availability.

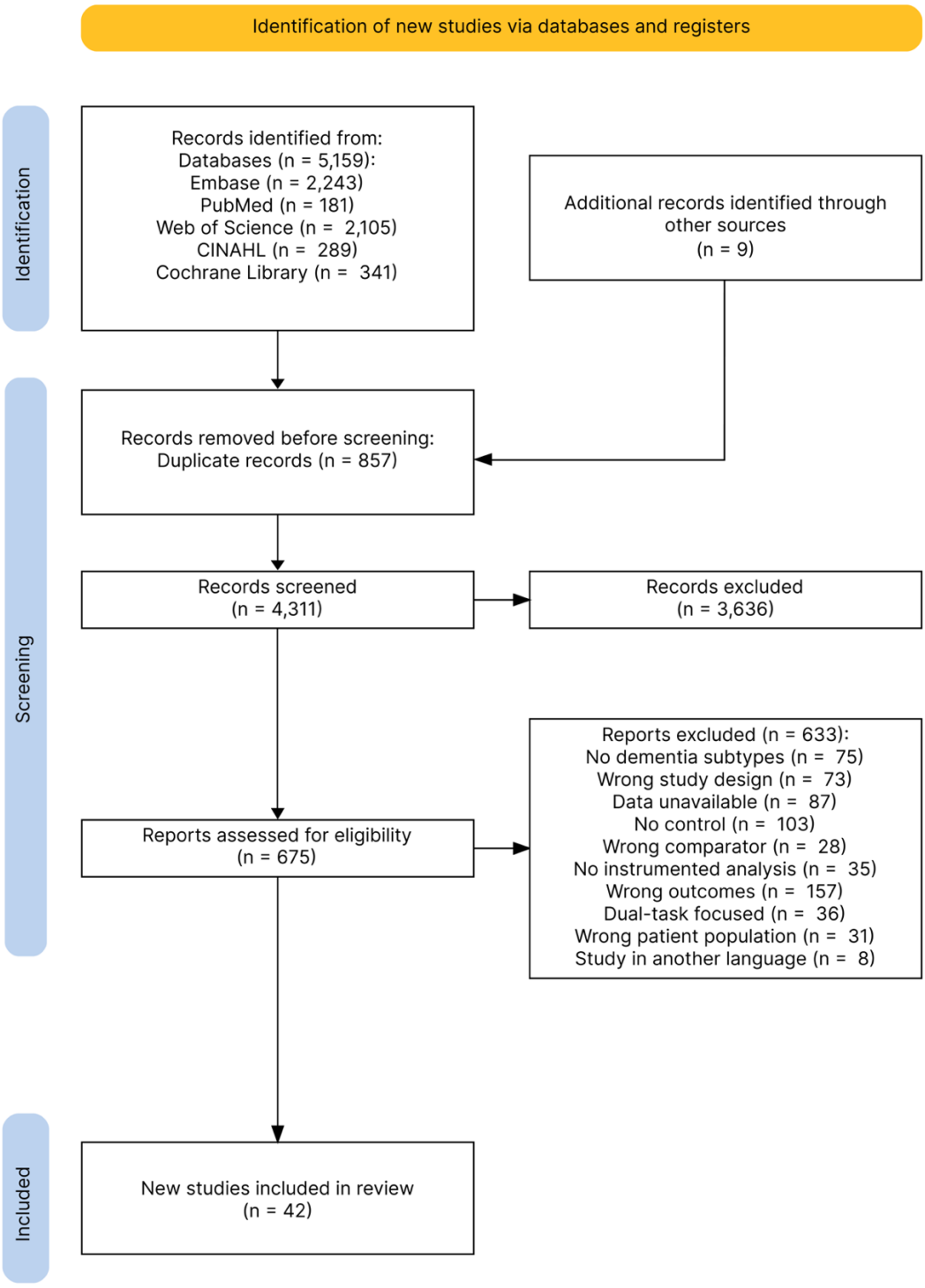
2.3 Results

2.3.1 Results of Database Search: Figure 2.1 illustrates the search results using a *PRISMA* flowchart. A total of 5,168 records were identified, with 5,159 from database searches and 9 from other sources. After removing 857 duplicate records, 4,311 studies remained for screening. Title and abstract screening led to the exclusion of 3,636 records. Among the 675 full-text articles assessed for eligibility, 633 were excluded based on predefined criteria, leaving 42 studies for inclusion in the final review. The primary reasons for exclusion were irrelevant study populations, lack of gait analysis data, and incomplete reporting of results. Our final review included 3,356 participants: 1,790 healthy controls, 1,264 individuals with Alzheimer's disease (AD), 186 with Lewy body dementia (LBD), 37 with frontotemporal dementia (FTD), and 265 with vascular dementia (VaD). The publication years of the studies ranged from 1983 to 2024, with a median year of 2015.

3.3.2 Summary of Participant and Study Characteristics: Figure 2.2 provides a heatmap summarizing the characteristics of the studies included in the analysis, categorized into participant and study characteristics. Of the 42 studies, AD was the most frequently studied dementia subtype (92.9%), while LBD (14.3%), FTD (7.1%), and VaD (9.5%) were less commonly represented. Dementia severity was inconsistently reported, but mild cases were more frequently included (28.6%). The majority of studies (81.0%) focused on participants aged 65-80 years. Although participant sex distribution and mean age (with standard deviations) were consistently reported, data on walking aid usage and self-selected walking speed were available in only about a quarter of the studies.

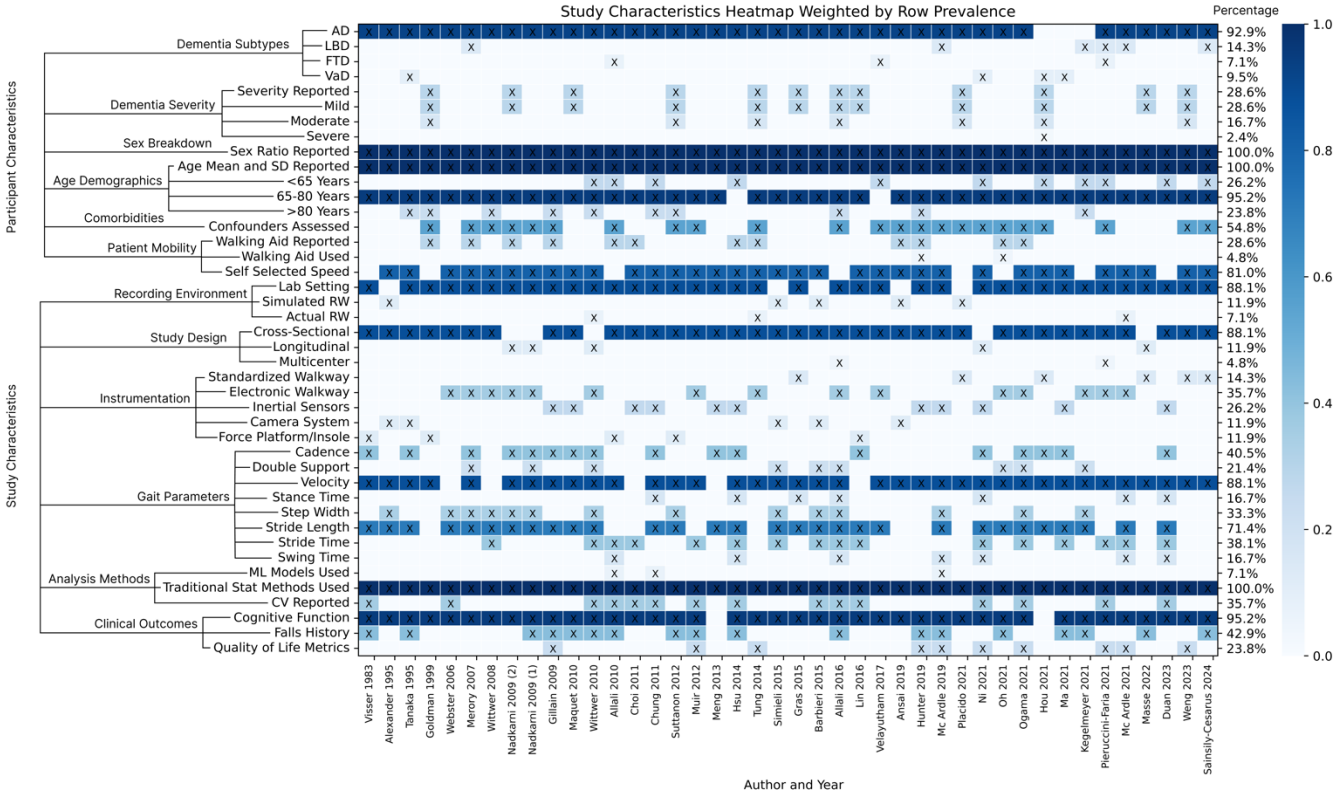
Regarding study characteristics, most investigations (88.1%) were conducted in laboratory settings, with fewer utilizing simulated (11.9%) or real-world environments (7.1%). Study designs varied, with 88.1% being cross-sectional, 11.9% longitudinal, and 4.8% multicenter studies. Instrumentation used included standardized walkways (14.3%), electronic walkways (35.7%), inertial sensors (26.2%), camera, and force systems, although the latter two were less frequently used. Gait parameters such as cadence (40.5%), gait velocity (88.1%), and stride length (71.4%) were commonly assessed, while parameters like step width and swing time appeared in fewer studies. Traditional statistical methods were employed in all studies, with machine learning models used in 7.1% of the analyses. Clinical outcomes were inconsistently reported; cognitive function was assessed in 95.2% of studies, falls history in 42.9%, and quality of life metrics, such as mental health and *Activities of Daily Living* (ADL) indices, in 23.8%. These inconsistencies

highlight the variability in study focus and the need for standardized reporting to better capture the broader impact of dementia on patients' lives.



Note: CENTRAL, Cochrane Central Register of Controlled Trials; CINAHL, Cumulative Index to Nursing and Allied Health Literature.

Figure 2.1 *PRISMA Flow Diagram of Study Selection Process*



Note: AD, Alzheimer's disease; LBD, Lewy body dementia; FTD, frontotemporal dementia; VaD, vascular dementia; RW, real-world; ML, machine learning; CV, coefficient of variation.

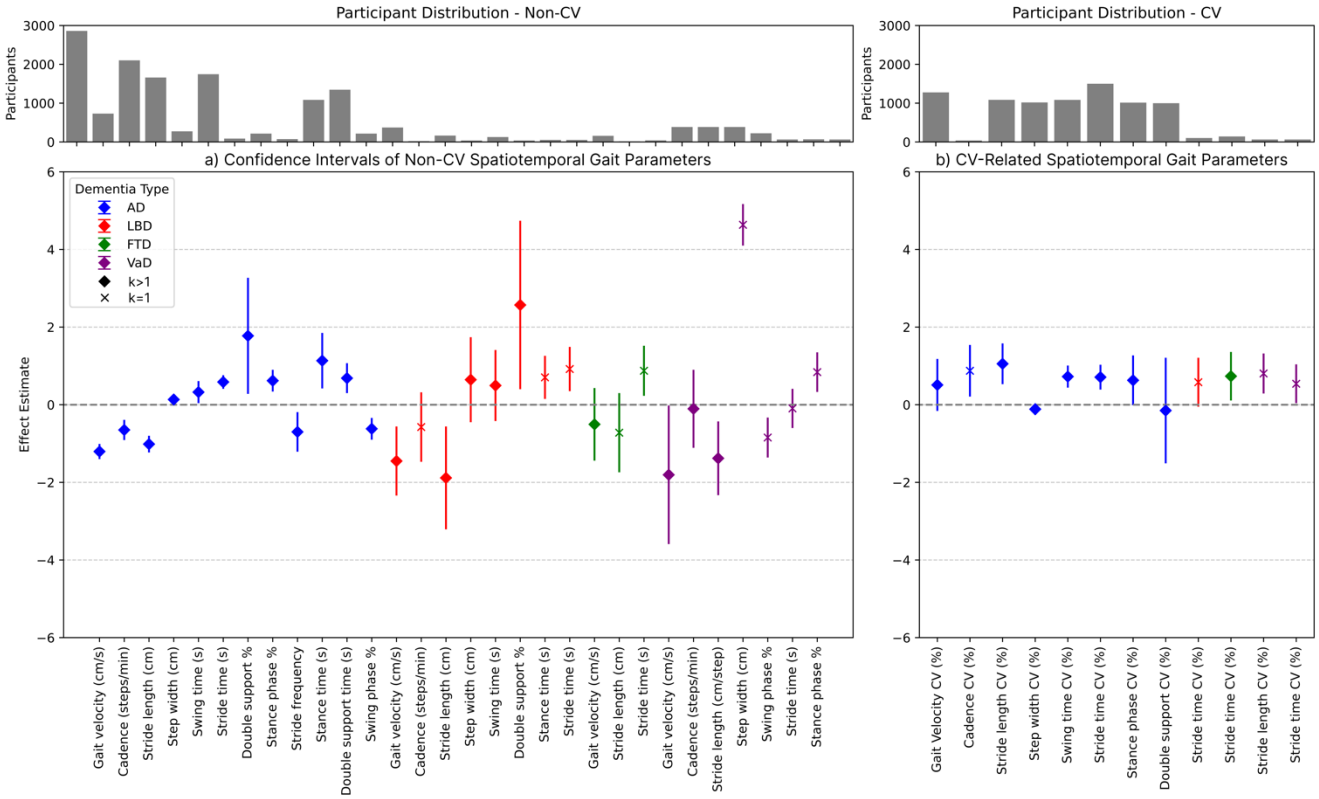
Figure 2.2 Summary of Study Characteristics and Reporting Trends

2.3.3 Comparative Gait Patterns Between Dementia Subtypes and Controls: Figure 2.3 summarizes the effect estimates and confidence intervals for various spatiotemporal gait parameters. In interpreting these results, it is important to consider that standardized mean differences (SMDs) represent deviations from healthy controls, with negative values indicating reductions in gait parameters (e.g., lower gait velocity or shorter stride length) and positive values indicating increases (e.g., prolonged stance time or greater step width). Larger absolute values reflect greater deviations from control participants, suggesting more pronounced impairment.

The results show that AD had the most robust reporting, with significant negative effect sizes observed for gait velocity (-1.20 [95% CI: -1.40, -1.01]), cadence (-0.65 [95% CI: -0.91, -0.39]), and stride length (-1.01 [95% CI: -1.23, -0.80]). LBD and VaD exhibited similar gait deficits in these parameters, but with wider confidence intervals and lower precision, suggesting a weaker evidence base.

Swing time (AD: 0.32 [95% CI: 0.04, 0.61], LBD: 0.50 [95% CI: -0.42, 1.41]), stance time (AD: 1.13 [95% CI: 0.42, 1.85], LBD: 0.71 [95% CI: 0.15, 1.26]), and the proportion of double support phase (AD: 0.68 [95% CI: 0.30, 1.07], LBD: 2.57 [95% CI: 0.40, 4.74]) showed significantly increased effect

sizes in both AD and LBD. Notably, the stance time result for LBD was based on a single study, highlighting the limited evidence and the need for cautious interpretation.



Note: The figure shows effect estimates with 95% confidence intervals for non-CV (a) and CV-related (b) gait parameters across dementia subtypes. Larger effect sizes indicate greater gait impairments compared to controls.

Figure 2.3 Effect Estimates of Spatiotemporal Gait Parameters Across Dementia Subtypes

Stride time showed significantly increased effect sizes in AD (0.58 [95% CI: 0.41, 0.76]), LBD (0.92 [95% CI: 0.35, 1.49]), and FTD (0.87 [95% CI: 0.23, 1.52]), but not in VaD (-0.10 [95% CI: -0.60, 0.41]). However, findings for LBD, FTD, and VaD were derived from a single study each, limiting the robustness of these results. Similarly, step width showed a significantly increased effect size only in VaD (4.63 [95% CI: 4.10, 5.17]), though this result was also supported by a single study, indicating the need for further validation.

For coefficient of variation (CV)-related gait parameters, AD showed the most robust reporting. Significant increases in stride length CV (1.06 [95% CI: 0.53, 1.58]), swing time CV (0.73 [95% CI: 0.44, 1.01]), and stride time CV (0.71 [95% CI: 0.39, 1.03]) were found. Stride time CV also showed significant increases in FTD (0.74 [95% CI: 0.11, 1.36]). The effect estimates for LBD (0.58 [95% CI: -0.05, 1.21]) and VaD (0.54 [95% CI: 0.04, 1.04]) were comparable, but both were based on data from a single study,

limiting the strength of the evidence. For full effect estimates and heterogeneity values, see Table 2.2 and Table 2.3.

Table 2.2 Summary Effect Sizes of Single-study Dementia Subtypes Spatiotemporal Parameters

Disorder	**k**	**n_D**	**n_C**	**SMD**	**p-value**	**95% CI**
Cadence[81],[82]						
$AD_{(CV)}$	1	19	19	0.88	0.01	[0.21, 1.54]
LBD	1	10	10	-0.57	0.03	[-1.47, 0.32]
Stance Time/Percent [83],[84]						
$LBD_{(s)}$	1	28	25	0.71	0.01	[0.15, 1.26]
$VaD_{(\%)}$	1	32	32	0.84	0.001	[0.33, 1.35]
Stride Length [85]						
FTD	1	8	8	-0.72	0.17	[-1.74, 0.30]
Stride Time [83],[86],[84]						
$LBD_{(s)}$	1	28	25	0.92	0.002	[0.35, 1.49]
$FTD_{(s)}$	1	19	22	0.87	0.008	[0.23, 1.52]
$VaD_{(CV)}$	1	32	32	0.54	0.04	[0.04, 1.04]
Swing Time/Percent [84]						
$VaD_{(\%)}$	1	32	32	-0.85	0.001	[-1.36, -0.33]

Note: Some spatiotemporal parameters in the "Disorder" column are denoted with (s) for absolute values, (%) for percentages, or (CV) for coefficients of variation. The selection of measurement type is based on the nature of the parameter being measured and the appropriate dimension (time or length). AD: Alzheimer's disease; LBD: Lewy body dementia; FTD: frontotemporal dementia; VaD: vascular dementia; k: number of studies; n_D: number of demented patients; n_C: number of control patients; SMD: standard mean difference; CI: 95% confidence interval.

Table 2.3 Summary Effect Sizes of Dementia Subtypes Spatiotemporal Parameters From Meta-analysis

Disorder	**k**	**n_D**	**n_C**	**SMD**	**p-value**	**95% CI**	**I^2**
Meta Analysis of Cadence $^{AD:}$ [81], [82], [84], [87], [88], [89], [90], [91], [92], [93], [94], [95]; $^{VaD:}$ [84], [94], [96], [97]							
AD	12	406	325	-0.65	<0.001	[-0.91, -0.39]	46%
VaD	4	121	265	-0.11	0.83	[-1.11, 0.90]	94%
Meta Analysis of Double Support Time/Percent $^{AD(s):}$ [91], [93], [98], [99], [100]; $^{AD(\%):}$ [81], [82], [101]; $^{AD(CV):}$ [81], [98], [101]; $^{LBD(\%):}$ [82], [102]							
$AD_{(s)}$	5	455	891	0.68	<0.001	[0.30, 1.07]	82%
$AD_{(\%)}$	3	44	44	1.50	0.10	[-0.28, 3.27]	91%
$AD_{(CV)}$	3	230	769	-0.15	0.83	[-1.51, 1.21]	95%
$LBD_{(\%)}$	2	18	20	2.57	0.02	[0.40, 4.74]	81%

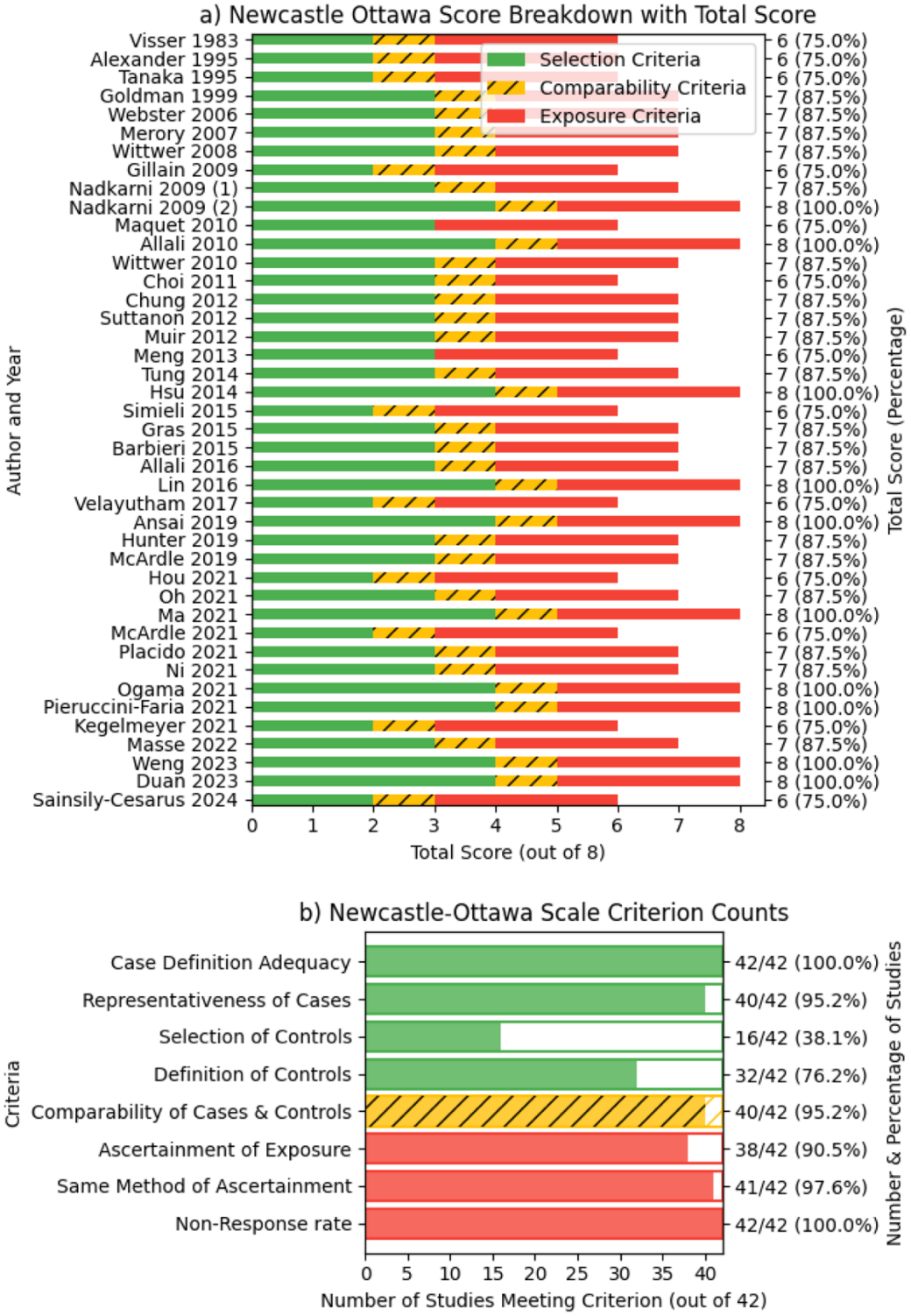
Meta Analysis of Gait Velocity [AD:] [75], [81], [82], [83], [84], [85], [86], [87], [88], [91], [92], [93], [94], [95], [98], [99], [100], [101], [103], [104], [105], [106], [107], [108], [109], [110], [111], [112], [113], [114], [115], [116], [117], [118]; [AD(CV):] [81], [93], [98], [101]; [LBD:] [75], [82], [102], [111], [116], [119]; [FTD:] [85], [86], [111]; [VaD:] [84], [94], [96], [97]							
AD	34	1212	1649	-1.20	<0.001	[-1.40, -1.01]	76%
$AD_{(CV)}$	4	417	860	0.51	0.14	[-0.16, 1.18]	94%
LBD	6	186	187	-1.45	0.001	[-2.34, -0.56]	89%
FTD	3	37	121	-0.51	0.29	[-1.44, 0.43]	77%
VaD	4	265	121	-1.80	0.05	[-3.59, -0.02]	97%
Meta Analysis of Stance Time/Percent [AD(s):] [83], [88], [98], [107]; [AD(%):] [84], [88], [95]; [AD(CV):] [87], [88], [98]							
$AD_{(s)}$	4	262	823	1.13	0.002	[0.42, 1.85]	88%
$AD_{(\%)}$	3	93	123	0.62	<0.001	[0.34, 0.90]	0%
$AD_{(CV)}$	3	226	788	0.63	0.05	[-0.01, 1.27]	77%
Meta Analysis of Step Width [AD:] [75], [81], [82], [91], [92], [93], [98], [99], [101], [103], [112], [120], [121]; [AD(CV):] [81], [98], [101], [120]; [LBD:] [82], [102]							
AD	13	614	1046	0.13	0.06	[0.00, 0.27]	14%
$AD_{(CV)}$	4	240	779	-0.11	0.13	[-0.26, 0.03]	0%
LBD	2	18	20	0.65	0.25	[-0.45, 1.74]	62%
Meta Analysis of Stride Length [AD:] [75], [81], [82], [83], [84], [85], [87], [88], [90], [91], [92], [93], [94], [95], [98], [99], [101], [103], [105], [107], [109], [112], [114], [120], [121]; [AD(CV):] [84], [89], [98], [101], [114], [120]; [LBD:] [75], [82], [83], [102]; [VaD:] [84], [94]							
AD	26	823	1280	-1.01	<0.001	[-1.23, -0.80]	68%
$AD_{(CV)}$	6	273	813	1.06	<0.001	[0.53, 1.59]	76%
LBD	4	91	74	-1.88	0.005	[-3.21, -0.56]	90%
VaD	4	265	121	-1.38	0.004	[-2.33, -0.43]	92%
Meta Analysis of Stride Time [AD(s):] [81], [83], [84], [86], [88], [89], [93], [95], [98], [99], [101], [110], [121], [122]; [AD(CV):] [81], [84], [86], [88], [89], [95], [98], [101], [110], [111], [122]; [FTD(CV):] [86], [111]							
$AD_{(s)}$	14	642	1104	0.58	<0.001	[0.41, 0.76]	42%
$AD_{(CV)}$	11	457	1043	0.71	<0.001	[0.39, 1.03]	79%
$FTD_{(CV)}$	2	29	113	0.74	0.02	[0.11, 1.36]	44%
Meta Analysis of Swing Time/Percent [AD(s):] [75], [83], [88], [95]; [AD(%):] [84], [88], [95]; [AD(CV):] [88], [95], [98]; [LBD(s):] [75], [83]							
$AD_{(s)}$	4	130	145	0.32	0.02	[0.04, 0.61]	24%
$AD_{(\%)}$	3	93	123	-0.62	<0.001	[-0.90, -0.34]	0%
$AD_{(CV)}$	3	258	826	0.73	<0.001	[0.44, 1.01]	48%
$LBD_{(s)}$	2	73	54	0.50	0.29	[-0.42, 1.41]	84%

Note: Some spatiotemporal parameters in the "Disorder" column are denoted with (s) for absolute values, (%) for percentages, or (CV) for coefficients of variation. The selection of measurement type is based on the nature of the parameter being measured and the appropriate dimension (time or length). AD: Alzheimer's disease; LBD: Lewy body dementia; FTD: frontotemporal dementia; VaD: vascular dementia; k: number of studies; n_D: number of demented patients; n_C: number of control patients; SMD: standard mean difference; CI: 95% confidence interval; I^2: heterogeneity statistic.

2.3.4 Risk of Bias Assessment Findings: Figure 2.4 presents the *NOS* assessment of study *ROB* across the included case-control studies, evaluating selection, comparability, and exposure criteria. The overall risk of bias was low, with a mean *NOS* score of 6.92 out of 8. The selection criteria had the most gaps, highlighting inconsistencies in case definitions, representativeness of cases, and control selection across studies. In contrast, most studies performed well in the comparability domain, showing adequate control for confounding factors, and in the exposure domain, which assessed exposure ascertainment and response rates. These findings emphasize the variability in study inclusion criteria and the importance of improving selection methods in future research.

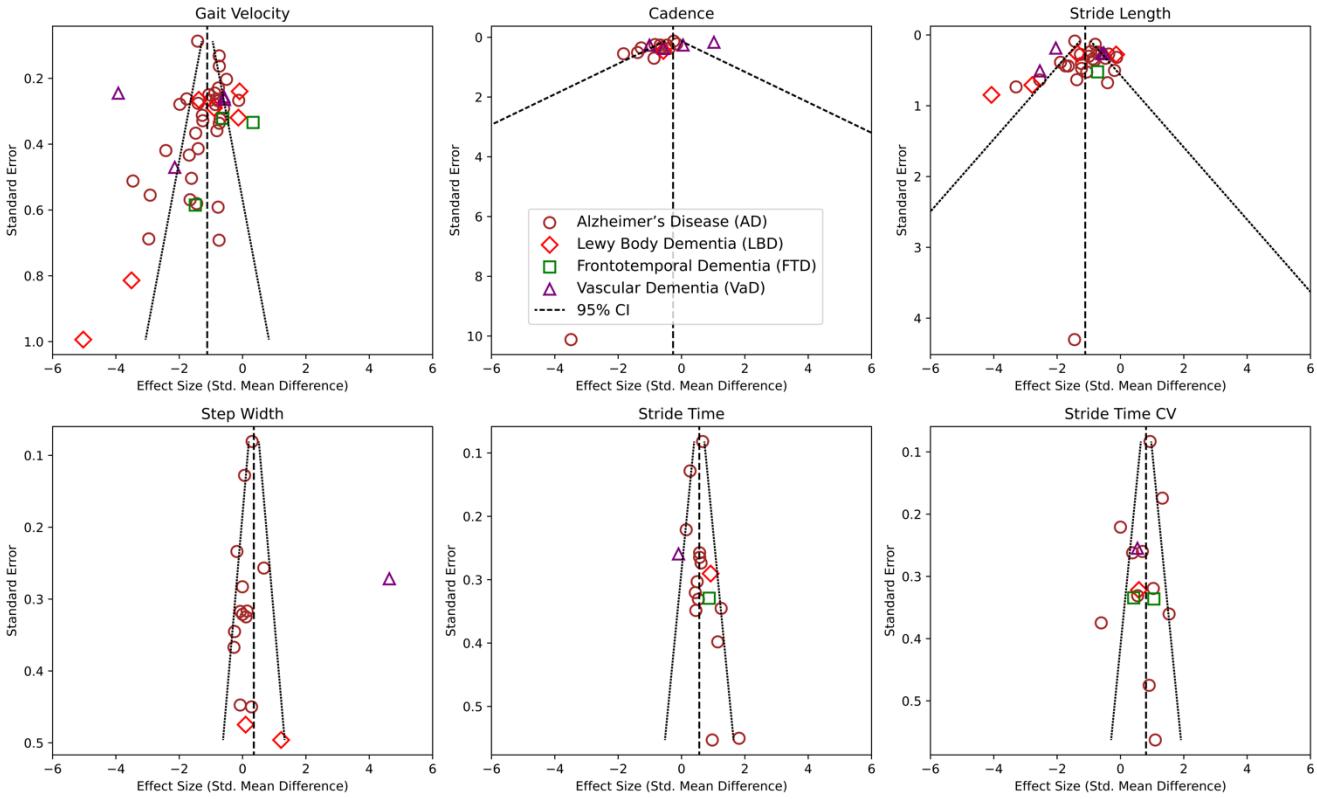
2.3.5 Assessment of Publication Bias: Figure 2.5 presents funnel plots assessing publication bias for six of the most frequently reported gait parameters: gait velocity, cadence, stride length, step width, stride time, and stride time CV. Stride length (-1.127, $P < 0.001$) and cadence (-0.506, $P = 0.007$) exhibited significant asymmetry, suggesting potential publication bias, likely due to missing studies with smaller effect sizes. Gait velocity (-0.019, $P = 0.947$) did not show statistically significant asymmetry, but it had a high number of barely significant studies (9 studies with p-values < 0.05), raising concerns of selective reporting. Conversely, step width (0.567, $P = 0.550$), stride time (-0.078, $P = 0.752$), and stride time CV (0.624, $P = 0.167$) displayed more symmetrical distributions, indicating minimal bias. The exclusion of other gait parameters was due to limited data and reporting inconsistencies. While most parameters appeared robust, the strong asymmetry and high Egger's significance in stride length and cadence warrant cautious interpretation. Funnel plots with fewer than 10 studies are illustrative, as small-sample effects can distort asymmetry. We excluded such outcomes to ensure reliability and avoid misleading interpretations.

2.3.6 Subgroup Analysis of Measurement Instrumentation: Our analysis, which adjusted for different measurement devices, revealed differences in gait parameters among dementia participants. Notably, electronic walkways detected increases in step width and swing time — findings not observed with camera-based systems or inertial sensors. The influence of measurement tools was particularly evident in the variability of stride length across dementia subtypes, especially in VaD. However, given the high heterogeneity and limited study representation within subgroups, numerical estimates were not emphasized, as isolated effect sizes may not reliably reflect broader trends. Instead, as discussed in the next section, we present a narrative synthesis to better capture methodological differences and highlight key insights. These findings underscore the need for standardized methodologies in future research to improve consistency in gait analysis.



Note: (a) Newcastle-Ottawa Scale (NOS) scores for each study, categorized by Selection (green), Comparability (yellow, hatched), and Exposure (red) criteria. (b) Criterion-level breakdown showing the number and percentage of studies meeting each NOS criterion.

Figure 2.4 Newcastle-Ottawa Scale Assessment of Study Risk of Bias



Note: Funnel plots show the standard error (SE) versus standardized mean difference (SMD) for six gait parameters across dementia subtypes. Subgroups: AD (brown circles), LBD (red diamonds), FTD (green squares), and VaD (purple triangles). Symmetry around the dashed line suggests minimal bias, while asymmetry may indicate potential bias or heterogeneity.

Figure 2.5 Funnel Plots Assessing Publication Bias in Gait Parameters

2.4 Discussion

2.4.1 Findings Summary: This systematic review and meta-analysis confirm the substantial gait discrepancies across different dementia subtypes, supporting prior research [119], [123]. Our findings emphasize the potential of gait analysis as a valuable diagnostic and prognostic tool in dementia research, with important implications for therapeutic interventions. The majority of the included studies focused on Alzheimer's disease (AD) (92.9%), with fewer studies examining Lewy body dementia (LBD), frontotemporal dementia (FTD), and vascular dementia (VaD), highlighting an imbalance in the available evidence across subtypes. Despite this, the overall risk of bias was low, with a mean *NOS* score of 6.92 out of 8, providing reasonable confidence in the results.

2.4.2 Alzheimer's Disease: In agreement with previous studies, our meta-analysis confirms that individuals with AD exhibit significant alterations in gait parameters, particularly in walking speed, cadence, and stride length. Although our analysis did not specifically emphasize prolonged stance time, existing literature suggests that this may act as a compensatory mechanism for postural instability in AD [70]. These

adaptations underscore the connection between locomotion and cognitive function in AD, pointing to potential therapeutic targets. Furthermore, our findings reveal significant variability in gait patterns within the AD population, potentially influenced by differences in study methodologies and participant characteristics.

2.4.3 Lewy Body Dementia: Our meta-analysis indicates that individuals with LBD may exhibit more pronounced gait deficits than those with AD, reflected in larger effect sizes. This aligns with previous studies highlighting the greater challenges faced by individuals with LBD [123], [124]. However, due to overlapping confidence intervals and high heterogeneity, direct statistical comparisons between subtypes should be interpreted cautiously. Rather than relying solely on significance testing, we emphasize the importance of effect size magnitude and consistency across studies.

The observed variations in stride length, cadence, and walking speed reflect the unique pathophysiology of LBD, while also raising questions about potential compensation strategies. Specifically, our data suggest that individuals with LBD may adopt a wider base of support, indicated by increased step width, to maintain postural stability while walking. However, the results section showed that step width was significantly increased only in VaD, and the findings for LBD were inconsistent across studies [82], [102] warranting cautious interpretation. The observed heterogeneity in step width results underscores the need for further research to determine the clinical relevance of this measure in LBD populations.

2.4.4 Frontotemporal Dementia: FTD, which typically presents at an earlier age [73], showed minimal gait disturbances in our meta-analysis. However, stride time was significantly increased in FTD, suggesting that while gait is generally preserved compared to other subtypes, certain gait parameters may still be affected. The preservation of most gait metrics despite cognitive decline suggests that FTD follows a different pathological pathway compared to other dementias. Autopsy studies have shown that FTD predominantly affects the frontal and temporal lobes, areas less involved in motor control, while AD affects regions such as the hippocampus and posterior parietal cortex, which are more directly associated with motor function and spatial navigation [125], [126]. Further research is needed to explore these anatomical differences and their functional implications for gait in dementia.

2.4.5 Vascular Dementia: Our meta-analysis confirms that VaD is associated with distinct gait disturbances, particularly increased stance time [84]. This aligns with prior studies that link VaD to frontal gait disorders [70], [127]. Our findings primarily highlighted alterations in cadence, gait velocity, and stride length, though a shuffling gait pattern or episodic freezing was not consistently reported. These observations should be interpreted with caution, as the complexity of gait disturbances in VaD and the limited data on specific gait phenomena suggest that further detailed research is needed. Future studies

should focus on better understanding the relationship between gait alterations in VaD and more intricate gait phenomena to uncover their underlying neuropathological mechanisms.

2.4.6 Instrument Choice and Subgroup Effects: The reliability of gait analysis can be influenced by the type of measurement instrument used. Stationary systems like force platforms and electronic walkways (e.g., the Kistler force platform [114] and the GAITRite system [75], [102], [111]) excel in controlled environments, accurately recording ground reaction forces and temporal gait parameters crucial for detailed gait cycle analysis. Our results indicated that stride length across dementia subtypes was more influenced by the choice of measurement device than by step width and swing time, which were previously thought to be more variable.

In contrast, optoelectronic [99], [101] and camera-based systems [89], [94], [104] specialize in spatial tracking, capturing movement from multiple angles to provide a comprehensive three-dimensional reconstruction of the gait cycle. These systems are invaluable for understanding spatial gait abnormalities. Additionally, wearable sensors [83], [88], [108] represent another valuable tool for gait analysis and the final category of instruments used. Our experience with inertial measurement units (IMUs) and the development of smartphone sensor applications for gait analysis have enriched our perspective on these devices [39], [40], [41], [42], [78]. Wearable sensors, such as accelerometers, are particularly useful for capturing real-world, naturalistic gait data, offering complementary insights to those gathered from stationary systems. The variability in results across different instruments, along with the need to balance their respective trade-offs, presents a unique challenge in gait analysis research. To address these challenges, harmonizing data across instruments and establishing standardized methodologies are essential next steps for advancing this field.

***2.4.7 Strengths and Weaknesses*:** A key strength of this study is the comprehensive inclusion of studies reporting spatiotemporal gait parameters across multiple dementia subtypes. The use of subgroup analysis enabled a more detailed comparison of gait characteristics, facilitating the identification of distinct patterns across subtypes.

However, several limitations must be acknowledged. First, the inclusion of only English-language publications may have introduced selection bias, potentially overlooking relevant studies in other languages. Second, the a priori exclusion of gray literature, including archives, preprints (e.g., arXiv, bioRxiv, medRxiv), and conference abstracts, may have led to the omission of relevant data, contributing to potential publication bias. Third, dementia severity was inconsistently reported across studies, with only 28.6% focusing on mild cases, limiting the generalizability of findings across different disease stages. Fourth, there was a significant imbalance in study representation, with Alzheimer's disease (AD) being far more extensively studied than Lewy body dementia (LBD), frontotemporal dementia (FTD), and vascular

dementia (VaD). As a result, findings for the less-represented subtypes should be interpreted with caution. Finally, while the overall risk of bias assessment suggested that most studies met methodological criteria, inconsistencies in study design and the limited availability of well-controlled research for certain subtypes highlight the need for more rigorous, standardized investigations.

***2.4.8 Implications for Practice, Policy, and Research*:** Gait analysis holds promise as a diagnostic tool for dementia, emphasizing the need for further exploration into the distinct gait characteristics of each dementia subtype and the role of demographic factors. To ensure more reliable results, future studies should aim for larger sample sizes, implement rigorous age and sex matching, and standardize methodologies. However, potential publication bias, particularly for cadence and stride length, as observed in our funnel plot analysis, should be considered when applying these results in clinical practice. As gait analysis becomes increasingly integrated into dementia diagnostics, it is crucial for research to carefully distinguish between symptoms and diagnostic criteria to avoid circular reasoning. This approach will be essential for developing unbiased and effective diagnostic tools in dementia care.

2.5 Conclusion

This meta-analysis of 42 studies and 3,356 participants confirms that distinctive spatiotemporal gait patterns can help differentiate dementia subtypes, including Alzheimer's disease (AD), Lewy body dementia (LBD), frontotemporal dementia (FTD), and vascular dementia (VaD). AD exhibited the most consistent gait deficits, with several gait parameters showing increased coefficient of variation (CV), reflecting greater gait inconsistency. However, data on CV measures were limited for other subtypes, highlighting the need for further investigation. Findings for other subtypes showed lower robustness and limited study representation. Despite methodological differences, gait analysis holds promise as a non-invasive diagnostic tool to support early detection and personalized care. However, inconsistencies in dementia severity reporting and study *ROB* underscore the need for standardized protocols and further high-quality research to improve its clinical applicability.

3. Large Language Models for Systematic Review Screening and Inclusion

The previous chapter identified how dementia research, like so many areas of healthcare, is inundated with heterogeneous data and study designs. This brings us to the pressing question: how can researchers cope with the escalating volume of scientific literature, ensure timely syntheses of current knowledge, and maintain rigor in their findings?

Chapter 3 introduces the first major computational pivot of this thesis: the development and evaluation of large language model (LLM) based approaches for accelerating systematic reviews. While Chapter 2 dealt with manually curated evidence and traditional meta-analytic methods, here we leverage recent advances in AI—particularly generative models like GPT-3.5 Turbo to automate labor-intensive aspects of literature screening and inclusion.

This chapter explores the feasibility, efficiency, and reliability of AI-assisted screening. In doing so, it responds directly to the shortcomings identified previously slow review cycles, potential reviewer fatigue, and the risk of missing critical studies. Chapter 3 establishes a new paradigm: one where emerging language technologies can significantly reduce the time and cost of evidence synthesis without entirely replacing human expertise. This sets the stage for further refinements explored in Chapter 4, where we experiment with more nuanced prompt engineering and threshold settings [*].

3.1 Introduction

3.1.1 Current Challenges of Conducting a Review: Systematic reviews and meta-analyses are cornerstone methodologies in evidence-based medicine, providing a comprehensive synthesis of research findings to inform clinical and policy decisions [128]. However, the traditional approach to conducting these reviews is labor-intensive and time-consuming, often requiring a year or more to complete and significant financial resources [62], [63], [129], [130]. The exponential growth in scientific publications further complicates the task, increasing both the complexity and the scope of reviews [64]. This scenario underscores a critical need for innovative methodologies that can streamline the review process without compromising its methodological rigor and accuracy.

The challenges of conducting systematic reviews extend beyond mere resource allocation. The inherent delay in incorporating the latest research findings into reviews due to publication lags adversely affects the timeliness and relevance of the synthesized evidence [17], [64], [131], [132]. Additionally, the

[*] In consideration for WIREs Data Mining and Knowledge Discovery (Impact Factor 6.4)

manual screening process, a key step in reviews, is not only time-consuming but also prone to inconsistencies and biases despite the expertise of reviewers [65], [133], [134]. The evolving landscape of systematic reviews, including rapid reviews and evidence synthesis for emergent health issues [64], further demands adaptive and efficient review processes that can cope with the dynamic nature of scientific research.

3.1.2 Text Mining and Automation in Systematic Reviews: The integration of text mining and automated technologies provides valuable tools for overcoming the challenges inherent in systematic reviews. Text mining, a field within data science, involves analyzing unstructured text to extract meaningful information and supports various stages of the review process such as study identification, screening, and data extraction—stages that are traditionally manual and labor-intensive [56], [57], [65], [134], [135], [136], [137]. These advancements have the potential to not only speed up the review process but also enhance the accuracy and objectivity of the data extracted, thereby improving the quality of systematic reviews.

Natural Language Processing (NLP), closely related to text mining, enables computers to understand and process human language, playing a crucial role in the automation of systematic reviews. It includes a range of tasks from information retrieval—where relevant articles are identified from large document collections—to document classification, which automates the inclusion or exclusion decisions in systematic reviews [64], [138]. With the advent of Large Language Models (LLMs) such as GPT [139], BERT [140], and their successors [139], [141], [142], [143], [144], [145], NLP has seen significant advancements, offering sophisticated capabilities for text analysis, and understanding that greatly enhance the review process [138]. These technologies allow for more nuanced and comprehensive analysis of scientific literature, leading to more sophisticated and scalable review methodologies.

3.1.3 Role and Impact of AI and LLMs: The integration of Artificial Intelligence (AI) and LLMs into the domain of systematic reviews marks a significant shift towards more efficient and effective evidence synthesis [146]. AI and NLP technologies automate the extraction and analysis of data from vast amounts of literature, streamlining the review process while maintaining, if not enhancing, the depth and breadth of analysis. The role of LLMs, characterized by their large parameter spaces and capacity for unsupervised learning, is particularly noteworthy. These models have demonstrated exceptional ability in understanding context, semantics, and the subtleties of language [147], [148], [149], [150], [151], [152], [153], [154], [155], [156], [157], [158], making them well-suited for tasks such as literature screening and data extraction in systematic reviews.

This study focuses on the application of LLMs in the article screening process. By automating this initial screening, AI tools allow researchers to dedicate more time to the complex tasks of data synthesis and interpretation. Moreover, AI-driven processes can potentially enhance consistency and reduce bias by standardizing the application of inclusion and exclusion criteria. As the field progresses, the role of AI and LLMs in systematic reviews is increasingly becoming a cornerstone for enabling more accessible, timely, and rigorous evidence synthesis, which is critical for informing healthcare policy and practice.

3.2 Methods

3.2.1 Overview of Original Reviews: Our research team recently conducted three separate reviews, each varying in degree and scope, and meticulously adhered to protocols established by standardized reporting committees [55], [57]. The first review investigated the relationship between dementia and spatiotemporal gait patterns to identify distinctive gait signatures. The second examined the latest advancements in cuffless blood-pressure monitoring devices, while the third focused on the impact of aging and comorbidities on long-COVID [159], [160], [161]. The selection process for each review involved pairs of authors screening titles, abstracts, and full texts, with any discrepancies resolved by a third author. Original inclusion and exclusion data were processed and gathered in .csv files, serving as our ground truth for the subsequent comparative analysis and validation depicted in our figures.

To assess whether these reviews could be accelerated using LLMs, we developed a pipeline to recreate and evaluate the screening process using generative AI. The following sections outline our reasoning behind model selection and cost analysis before detailing how the screening was operationalized.

3.2.2 Model Selection and Cost Analysis: Since the launch of ChatGPT in November 2022, the landscape of AI and language models has rapidly evolved, offering a range of options for various applications, including systematic reviews. During the planning phase of this study, back-of-the-envelope calculations identified GPT-3.5 Turbo as the most economical model, costing approximately $0.31 USD per 1,000 studies for a single pass of title-abstract screening. In contrast, GPT-4 Turbo was estimated to cost about 21 times more, at $6.61 USD per 1,000 studies. The newest model, GPT-4o, released in May 2024, is expected to cost $3.31 USD per 1,000 studies—about 50% less than GPT-4 Turbo—making it a promising candidate for future analysis.

3.2.3 Recreating Title-Abstract Screening with AI Models: Utilizing the LangChain framework [162], the title and abstract screening process was replicated using the OpenAI model GPT-3.5-turbo-0125,

employing Retrieval Augmented Generation (RAG) to accurately reflect the inclusion and exclusion criteria of the original human reviewers. In this RAG setup, the application provided a user interface with a text box where reviewers could enter any prompt. This feature allowed the reviewer to design a single prompt that was applied consistently across all titles and abstracts. For our screening protocol, the model was fed prompts based on the inclusion criteria of each review, designed to elicit binary 'true' or 'false' responses, simulating the decision-making framework of systematic reviews. The responses, along with associated metadata, were captured, cleaned for accuracy, and then exported as .csv files.

3.2.4 Recreating Full-Text Screening with AI Models: For the full-text screening, we adapted our approach to accommodate the OpenAI API's token limit of 4,096 tokens, roughly equivalent to 3,000 words. This constraint defines how much text and prompt content can be processed in a single API call. To work around this, we split each full-text article into smaller overlapping segments (about 750 words each) and stored them in a searchable vector database, Hierarchical Navigable Small World (HNSW) using text embeddings. This setup employed the OpenAI model, text-embedding-ada-002-v2, functioning similarly to a search engine, that uses cosine similarity to rank all the generated text snippets based on their relevance to a user-generated query.

In the full-text RAG setup, the same inclusion prompts designed to elicit binary ‘true’ and ‘false’ responses were used. Both the prompt and relevant text snippets were combined to form the input for GPT-3.5-turbo-0125. We configured the system to return three relevant snippets per query to effectively manage the constraints imposed by the API's token limit. This strategy leaves sufficient space (approximately 1,096 tokens) for users to employ complex query templates designed for either study screening or data extraction purposes.

3.2.5 Comparison Criteria and Discrepancy Analysis: To assess the effectiveness of GPT-3.5 Turbo's screening against the original human reviewers, we employed confusion matrices for both the title-abstract and full-text screening phases. Our comparison utilized key performance metrics such as accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), F1-score, and the Matthews Correlation Coefficient (MCC).

Discrepancy analysis was conducted by first identifying all true positives, true negatives, and false positives through a comparison of GPT-3.5 Turbo’s decisions against reviewer judgments during the title-abstract screening phase and securing the full texts for these articles. The analysis focused specifically on the subset of articles that were mutually recognized as 'included' by both the model and the reviewers at the title-abstract phase, ensuring that our evaluation of the model's full-text screening was

grounded in a directly comparable set of studies. Additionally, we assessed the consistency of the AI model by examining the remaining articles that GPT-3.5 Turbo initially classified as 'included' during the title-abstract phase but were later 'excluded' or continued to be 'included' in the full-text stage.

3.2.6 Evaluation of ChatGPT-3.5 Turbo Performance and Validation Procedure: To assess the performance of the LLM against a standardized control, we established a baseline using random binary classification. In this control setup, studies were arbitrarily classified as 'included' or 'excluded,' mimicking the binary decision-making process typical in systematic reviews. Alongside this, we conducted a self-validation test with two prompts to assess internal consistency within the LLM's responses, comparing agreement between the model's decisions when operating in explanation mode versus non-explanation mode.

To further evaluate our model, we adopted a permutation test similar to K-Folds Cross Validation. This involved shuffling the data and dividing both the model-generated dataset and the control dataset into K equal subsets. We computed standard performance metrics for each partition and then averaged the results across all partitions. Finally, to assess the statistical differences between the model and control, we conducted a One-way Analysis of Variance (ANOVA) comparing the mean performance metrics across the subsets and screening phases, using an alpha level of 0.05 to determine significance.

3.2.7 Data Handling and Ethical Considerations: Our research uses data from studies with established ethical approvals. Full prompt templates are provided as supplementary material. Additional data is available upon request.

3.3 Results

3.3.1 Outcomes of the Original Review: In Figure 3.1, we provide a comprehensive summary of the screening process facilitated by GPT-3.5 Turbo for each of the three reviews. This figure includes the number of studies identified, screened, included, and excluded, offering a direct comparison and overview of the model's performance in the review process.

3.3.2 Title-Abstract Confusion Matrices for each Review: Figure 3.2 showcases separate confusion matrices for the title-abstract inclusion task. These matrices illustrate the true positives, false positives, true negatives, and false negatives for GPT-3.5 Turbo in comparison to the original reviewers categorizations.

3.3.3 Title-Abstract Performance Metrics for each Review: Table 3.1 displays the performance metrics of GPT-3.5 Turbo’s responses for each review. This table showcases the model's accuracy, sensitivity,

specificity, predictive values, F1 score, and MCC in comparison to the original human reviewers. It also highlights that GPT-3.5 Turbo significantly outperformed the random classification control in all but two performance metrics (sensitivity: p-value = 0.22; F1-score: p-value = 0.21) emphasizing its effectiveness in accurately identifying relevant studies.

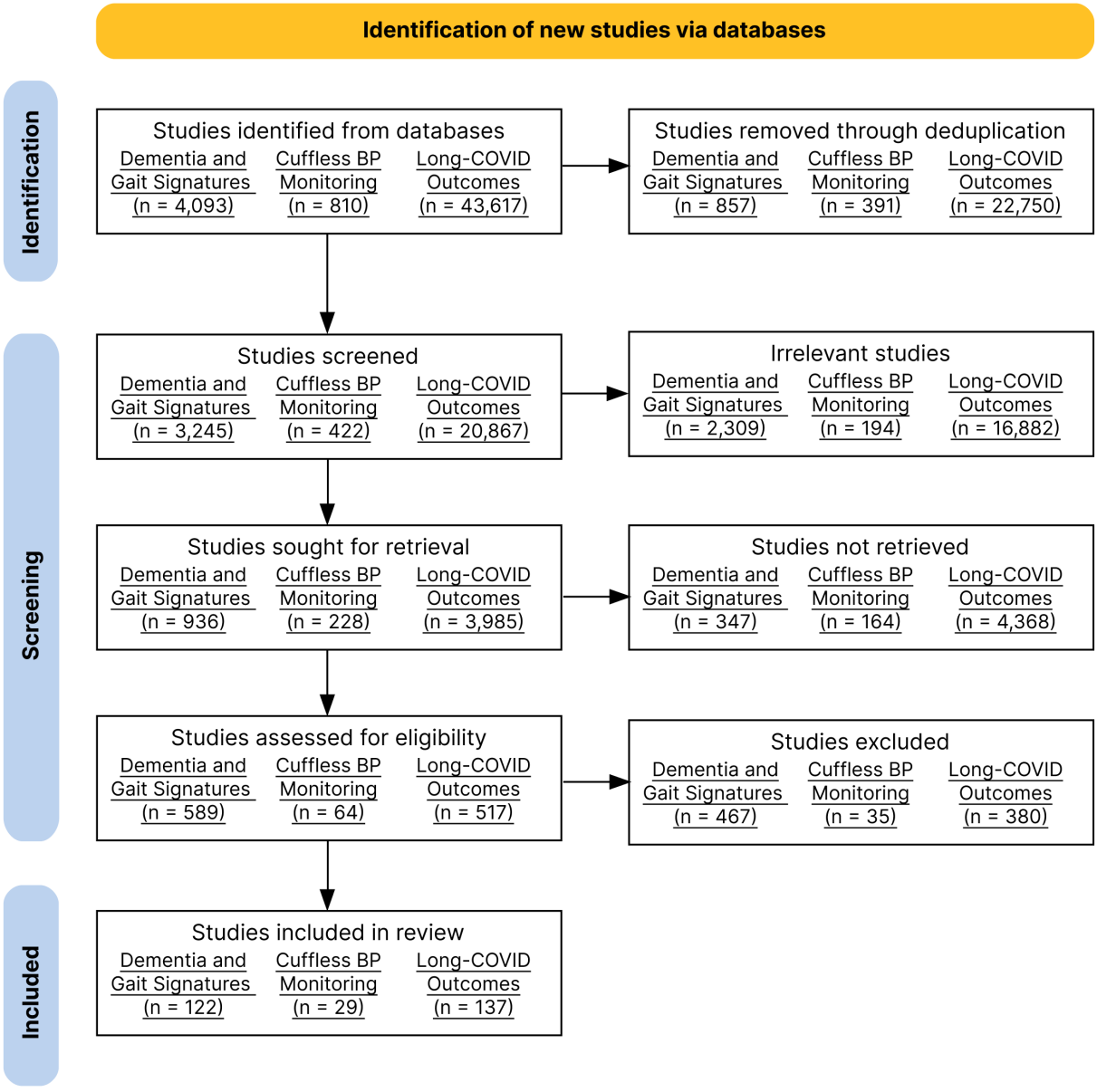
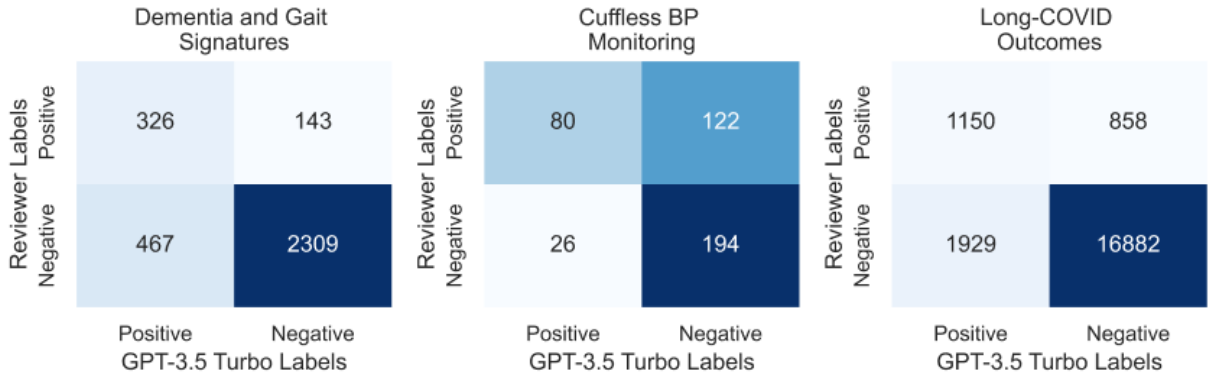


Figure 3.1 PRISMA Diagram of GPT-3.5 Turbo Screening Results Across Review Topics

3.3.4 Title-Abstract Self-Validation of AI Model Responses: Figure 3.3 presents confusion matrices for the AI model's responses, comparing when explanations were provided to when they were not. The results show a high level of agreement between the two sets of responses, although not absolute. Notably, when the model's decisions were matched with those of human reviewers, asking the model for an explanation

resulted in a slight drop in performance. This indicates that requesting explanations may subtly affect the model's output, suggesting the need for further investigation into how such prompts might alter the precision of AI-driven selections.



Note: Confusion Matrices Comparing GPT-3.5 Turbo Decisions Against Reviewer Judgments for Title-Abstract Screening Tasks in Dementia Gait, Cuffless Blood Pressure Monitoring, and Long-COVID Studies.

Figure 3.2 Alignment of Model Predictions with Human Reviewers

3.3.5 Confusion Matrices for Full-Text Screening: Figure 3.4 presents the data which compares the AI's decisions against human reviewers' judgments. We screened a total of 589 full-text articles on dementia and gait signatures, 64 on cuffless blood pressure monitoring, and 517 on Long-COVID outcomes. The upper section illustrates the alignment of full-text decisions for articles included by both the model and the reviewers in initial phases and the lower section highlights discrepancies for initially included studies.

Table 3.1 Comparative Analysis of GPT-3.5 Turbo's Performance Metrics for Title-Abstract Inclusion Across Review Topics

Metric	Dementia and Gait Signatures (10-fold)		Cuffless BP Monitoring (10-fold)		Long-Covid Outcomes (10-fold)	
	Control	Model	Control	Model	Control	Model
Accuracy	0.50 ± 0.03	**0.81 ± 0.02**	0.46 ± 0.09	**0.65 ± 0.08**	0.50 ± 0.01	**0.87 ± 0.01**
Sensitivity	0.49 ± 0.09	**0.69 ± 0.06**	0.46 ± 0.12	0.40 ± 0.12	0.50 ± 0.02	**0.57 ± 0.03**
Specificity	0.50 ± 0.02	**0.83 ± 0.03**	0.44 ± 0.10	**0.88 ± 0.05**	0.50 ± 0.01	**0.90 ± 0.01**
PPV	0.14 ± 0.02	**0.41 ± 0.05**	0.43 ± 0.14	**0.75 ± 0.11**	0.10 ± 0.01	**0.37 ± 0.04**
NPV	0.85 ± 0.03	**0.94 ± 0.01**	0.48 ± 0.10	**0.62 ± 0.09**	0.90 ± 0.01	**0.95 ± 0.00**
F1 Score	0.22 ± 0.04	**0.51 ± 0.05**	0.44 ± 0.13	0.51 ± 0.11	0.16 ± 0.02	**0.45 ± 0.04**
MCC	-0.01 ± 0.07	**0.43 ± 0.05**	-0.10 ± 0.17	**0.32 ± 0.14**	0.00 ± 0.01	**0.39 ± 0.04**

Note: Metrics were derived using a pseudo-K-folds cross-validation method. BP – Blood Pressure, PPV – Positive Predictive Value, NPV – Negative Predictive Value, MCC – Matthews Correlation Coefficient. Bolded values indicate a p-value < .05.

3.3.6 Performance Metrics for Full-Text Screening: Table 3.2 presents the performance metrics of the full-text screening protocol employed by GPT-3.5 Turbo. Sensitivity and MCC improved the most when model screening was applied. Notably, the model had significantly greater sensitivity in the full text screening phase compared to title and abstract screening, indicating a potential reduction in false negatives as more article content is assessed. While the model generally outperformed the control across metrics, many of these improvements did not reach statistical significance possibly due to the smaller full-text sample size.

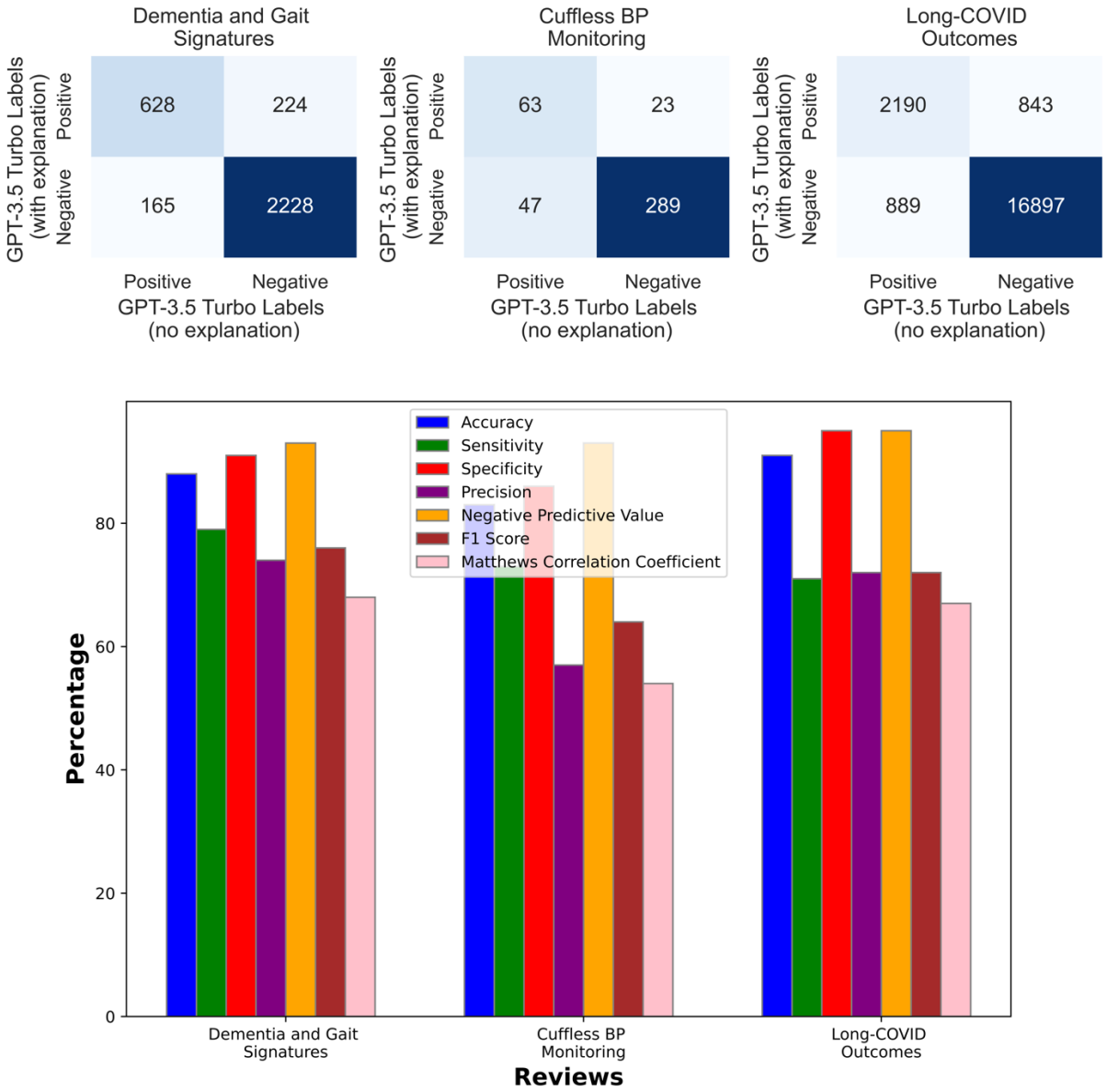
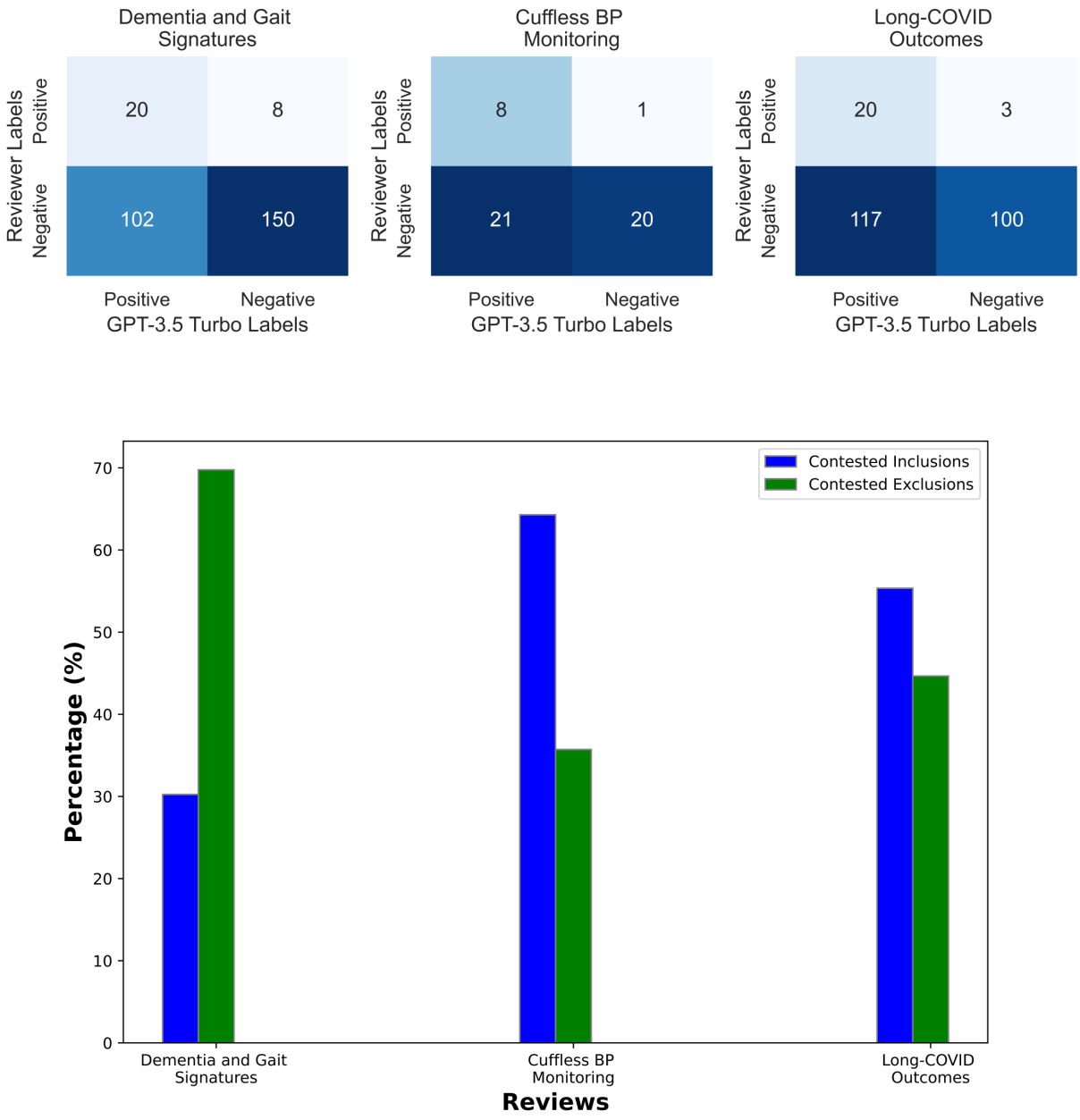


Figure 3.3 Comparative Analysis of GPT-3.5 Turbo Model Performance Across Reviews



Note: The upper panel presents confusion matrices comparing the AI model's positive and negative labels against reviewers' labels. The lower panel (N = 291, N = 14 and N = 262 respectively) shows the percentage of contested decisions differentiating between contested inclusions (blue) and contested exclusions (green).

Figure 3.4 Analysis of Full-Text Inclusion Task and Contested Decisions in AI-Assisted Review Screening

Table 3.2 Comparative Analysis of GPT-3.5 Turbo's Performance Metrics for Full-Text Inclusion Across Review Topics

Metric	Dementia and Gait Signatures (10-fold)		Cuffless BP Monitoring (5-fold)		Long-Covid Outcomes (10-fold)	
	Control	**Model**	**Control**	**Model**	**Control**	**Model**
Accuracy	0.53 ± 0.09	0.62 ± 0.10	0.41 ± 0.17	0.56 ± 0.21	0.47 ± 0.12	0.49 ± 0.07
Sensitivity	0.69 ± 0.27	0.69 ± 0.34	0.37 ± 0.42	**0.95 ± 0.11**	0.43 ± 0.42	**0.89 ± 0.18**
Specificity	0.52 ± 0.10	0.61 ± 0.12	0.40 ± 0.17	0.49 ± 0.24	0.47 ± 0.11	0.45 ± 0.07
PPV	0.14 ± 0.09	0.18 ± 0.12	0.14 ± 0.14	0.30 ± 0.15	0.08 ± 0.09	0.15 ± 0.08
NPV	0.92 ± 0.09	0.95 ± 0.04	0.77 ± 0.19	0.95 ± 0.11	0.87 ± 0.09	**0.97 ± 0.05**
F1 Score	0.22 ± 0.11	0.27 ± 0.16	0.20 ± 0.21	0.44 ± 0.17	0.14 ± 0.14	0.24 ± 0.12
MCC	0.10 ± 0.16	0.20 ± 0.18	-0.16 ± 0.37	**0.33 ± 0.22**	-0.07 ± 0.28	**0.20 ± 0.12**

Note: Metrics were derived using a pseudo-K-folds cross-validation method. BP – Blood Pressure, PPV – Positive Predictive Value, NPV – Negative Predictive Value, MCC – Matthews Correlation Coefficient. Bolded values indicate a p-value < .05.

3.4 Discussion

3.4.1 Role and Impact of AI and LLMs: The advent of AI and LLMs, epitomized by our LLM screening tool, marks a notable evolution in research methodologies for systematic reviews. LLMs such as BERT [140], Megatron-ML [145], GPT-3 [139], GPT-4 [143], and PaLM 2 [144], characterized by their large parameter sets and transformer architecture, have set new benchmarks in performance, illustrating the rapid evolution and potential of these models [64], [142]. Our protocol facilitates a paradigm shift, enabling more efficient screening and precise extraction of information from the expansive realm of scientific literature. This advancement not only speeds up the discovery process but also enriches the insights gained from data [146], aligning with the growing need for swift review methodologies amid the surge of preprint repositories [64].

Further enriching our understanding, our model's self-validation study offers a glimpse into the decision-making capabilities of AI. When comparing the model's internally consistent responses, whether reasoned or not, against human reviewers' decisions, a slight dip in performance was observed when the model explained its choices. This intriguing outcome hints at a complex, intuitive-like decision-making process within AI, akin to human cognition but distinct in its execution. These findings underscore the sophisticated nature of LLMs in systematic review processes and underscore the necessity for ongoing research to fully grasp AI's potential in enhancing both the efficiency and depth of research synthesis.

3.4.2 Integration with Existing Methodological Innovations and Tools: The introduction of LLM screening into the ecosystem of systematic review tools such as PICO Portal, DistillerSR, Covidence, and Rayyan exemplifies a leap forward in the automation of review workflows [60], [163], [164]. These tools, alongside innovations like RobotReviewer [165], TrialStreamer [166], and Abstrackr [167], have showcased their capability in extracting and evaluating information from scientific articles, thus aiding in judging study quality, and inferring treatment effects.

Unlike traditional human review processes, which can be subject to subconscious biases and sometimes lack expert knowledge, LLMs provide a consistent and replicable framework for decision-making and have demonstrated proficiency across various fields. However, our preliminary findings suggest that the wording of prompts can significantly impact the representation of articles included by the model, as illustrated by self-validation tasks and performance metrics of similar prompts (see eFigure 1). As other studies have indicated, integrating such AI technologies poses challenges, particularly in balancing effective filtering with accurate identification of pertinent studies [168]. This delicate equilibrium is crucial as we further incorporate AI into systematic reviews, ensuring that we harness both AI capabilities and human expertise without compromising the integrity and depth of research synthesis.

To contextualize the model's efficiency and effectiveness compared to traditional methods, it's important to note that manual review typically requires locking in strict inclusion/exclusion criteria early on, since retraining reviewers or adjusting guidelines mid-process can introduce bias and inconsistencies. By contrast, LLM-based workflows allowed us to rapidly iterate on prompt phrasing and decision logic without restarting or retraining a team, effectively letting us "re-review" thousands of papers under new criteria in a matter of hours. This flexibility not only reduced workload and fatigue but also made it easier to explore edge cases and refine judgment without derailing progress.

3.4.3 Cautious Reliance on Automated Screening Systems: The deployment of open-source frameworks like LangChain [162] have demonstrated the transformative potential of AI and LLMs in enhancing various tasks, including the efficiency, accuracy, and overall workload involved in systematic reviews. By leveraging automation for the initial screening of papers, the tool has markedly reduced the time and manual labor traditionally required. Yet, amidst these advancements, we acknowledge the inherent limitations of AI systems in detecting nuanced or edge-case studies, a domain where human reviewers' judgment and inclusivity play a crucial role.

The inclination of human reviewers to err on the side of inclusivity during the initial screening phases, adopting a 'better safe than sorry' approach to minimize the risk of overlooking potentially

relevant studies, highlights a critical area where AI models may falter. GPT-3.5 Turbo, while adept at streamlining the review process, has exposed a vulnerability in terms of false negatives, underscoring the technology's current limitations in fully grasping the subtleties of the full review procedure. This observation serves as a reminder of the need for cautious integration of AI-based screening systems within the systematic review workflow.

3.4.4 Data Validation and its Impact on Model Performance: Our evaluations have demonstrated GPT-3.5 Turbo's effectiveness in managing heterogeneous datasets and identifying opportunities for iterative enhancement. Recent studies have highlighted the susceptibility of LLMs to the "butterfly effect," where slight variations in input can precipitate significantly different outputs [169]. This sensitivity resonates with our findings from the self-validation tests, where subtle changes—such as requesting an explanation for decisions—markedly affected the model's alignment between trials.

Given these findings, a critical focus of our analysis was the examination of data integrity and its influence on the AI model's output. We observed minor inconsistencies in the reviewer-generated data, such as mismatches in titles, abstracts, and DOIs, which could potentially skew performance metrics. Manual restoration of these discrepancies had a negligible impact on the overall performance results, suggesting that the screening's robustness extends to accommodating minor data inconsistencies.

3.4.5 Impact on Policy and Practice: While the adoption of AI, particularly LLMs like GPT-3.5 Turbo, offers transformative potential for systematic review methodologies, it introduces new vulnerabilities that warrant careful consideration. Central among these concerns is the danger posed by reliance on a single AI model for all decision-making processes within reviews. The risks of such a dependency include vulnerability to adversarial attacks, where manipulated inputs could lead to inaccurate outcomes, and the presence of inherent biases within the AI model that could skew results in subtle yet significant ways.

Addressing these risks involves not only technological solutions but also a broader reconsideration of how AI tools are integrated into the review ecosystem. It necessitates a balanced approach that leverages the strengths of AI for efficiency and scalability while maintaining a critical awareness of its limitations and potential pitfalls. As we advance, fostering a diversified toolkit of AI models and open-source methodologies will be paramount in mitigating the risks associated with overreliance on any single system. This strategy will enhance the resilience of the review process, ensuring that it remains robust, transparent, and adaptable to the evolving landscape of scientific inquiry.

During data collection, we encountered several practical challenges that underscore the current barriers to automating systematic review processes. Our study screening protocol was constrained by

technical limitations, primarily due to reliance on textual inputs and difficulties handling multi-modal data, such as tables and figures. Additionally, even with access to institutional libraries, we faced obstacles in scraping for DOIs and accessing all relevant articles, hampered by publisher restrictions. These challenges highlight broader issues in the accessibility of scientific literature and point to the urgent need for infrastructural improvements. Such enhancements are crucial to support the seamless integration of AI tools in research synthesis and to ensure their effective utilization.

3.4.6 Future Directions: As we continue to refine our novel screening protocol, striking a balance between the speed and thoroughness of the review process remains a central challenge. Efforts are currently directed towards prompt engineering and comparing various models to minimize the exclusion of relevant literature. Future research should focus on enhancing the system's ability to process multi-modal inputs and expanding its capabilities for comprehensive, end-to-end review automation. Such advancements could optimize research synthesis and enable LLMs to assimilate findings across multiple disciplines. This could lead to unique and innovative insights, potentially revolutionizing the processing and utilization of complex, interdisciplinary information.

As it stands, while we advocate for the use of AI in systematic reviews for its undeniable benefits, we also emphasize the importance of integrating these technologies judiciously, ensuring that they complement rather than replace the nuanced judgment of human reviewers. This approach aims to harness the strengths of both AI and human expertise, optimizing the systematic review process without compromising the integrity and depth of research synthesis.

3.5 Conclusion

The development and application of our LLM screening protocol signifies a milestone in research methodology. In our comprehensive work involving the screening and validation of GPT-3.5 Turbo across three review topics and 24,534 studies, the model has demonstrated its potential as a sophisticated, efficient, and reliable model for systematic reviews and meta-analyses automation. Its ongoing evolution and refinement are paramount to keep pace with the rapid progression of scientific and technological innovations.

4. Few-Shot Systematic Review Inclusion with GPT and Scalable Prompt Engineering

Chapter 3 introduced the promise of LLMs as helpful assistants in the systematic review process, demonstrating that AI can indeed reduce workloads and improve efficiency. However, as we begin to rely on these models heavily, questions about precision, sensitivity, and the inherent trade-offs come to the forefront. How do we ensure that by increasing efficiency, we do not inadvertently exclude critical studies or introduce biases? How can we tune these advanced models to better balance the inclusion and exclusion of research?

Chapter 4 takes a deeper dive into these issues by examining the concept of ordinal prompting. This approach builds directly on the foundational work from Chapter 3, where we first deployed LLMs in the systematic review workflow. Now, we shift from a binary viewpoint—simply including or excluding an article, toward a more granular method, assigning confidence scores and using these to fine-tune our decisions.

By the end of Chapter 4, we will have a more sophisticated toolkit for harnessing AI in evidence synthesis, one that can adapt to different contexts and requirements. This methodological refinement is critical, as it sets the stage for tackling data-intensive problems outside the strict confines of literature reviews [*].

4.1 Introduction

4.1.1 Traditional Systematic Review Processes and Recent Advances in Artificial Intelligence:

Systematic reviews are essential to evidence-based practice [135], but they face considerable challenges due to their intensive labor requirements and extended timelines [62], [63], with costs potentially reaching up to $150,000 USD in labor alone [170]. These reviews involve multiple stages—including defining the research question, developing a protocol, conducting literature searches, screening titles and abstracts, full-text review, data extraction, and finally, data synthesis and reporting [55], [56], [57]—each of which requires meticulous effort and is vulnerable to human error and bias [171], [172], [173].

Completing a systematic review can be a lengthy process; a survey of 195 registered reviews found that the average review took 67.3 weeks to complete, with some extending as long as 186 weeks [63]. Moreover, despite the Cochrane Collaboration's recommendation to update reviews every two

[*] In consideration for the Journal of the American Medical Informatics Association (Impact Factor 7.9)

years, many remain outdated, limiting their relevance to current practice [174]. As scientific literature continues to grow exponentially, these traditional methods increasingly struggle with data saturation and reviewer fatigue, both of which threaten the timeliness and applicability of findings.

While a 2014 study identified fifteen stages in the systematic review process with potential for automation, significant advancements have been made over the past decade to realize this potential [175]. In 2024, Artificial Intelligence (AI) driven text mining techniques now allow for the transformation of unstructured text into machine-readable formats compatible with machine learning (ML) models, establishing a solid foundation for more efficient and effective automation in systematic reviews.

4.1.2 Text Mining Techniques and Tools in Systematic Review Automation: Before text data can be used with ML algorithms, it must first be converted into suitable formats. This begins with tokenization, where text is broken into smaller units or "tokens," such as words, characters, or groups of consecutive words (N-grams). A recent 2024 scoping review [176] identified 123 studies focused on automation technology for systematic reviews. N-grams were a popular method within the bag-of-words framework for capturing local context within text sequences, which counts token occurrences without regard to order [177].

Beyond bag-of-words, advanced feature extraction methods included word and sentence embeddings [178], [179], term frequency-inverse document frequency (TF-IDF) [180], and Latent Dirichlet Allocation (LDA) [181]. Embeddings capture semantic relationships in high-dimensional space, as shown by [182], [183], [184]. TF-IDF assigns weights to words by balancing document frequency with corpus-wide prevalence, allowing researchers to highlight distinctive terms for classification [185], [186]. Meanwhile, LDA groups words into topics based on co-occurrence patterns, helping to identify thematic structures within texts [187], [188], [189].

After transforming text into feature representations, ML models classify documents based on these features. Support Vector Machines (SVM) [190], for instance, handle high-dimensional, sparse datasets, and remain a popular choice for binary tasks like inclusion and exclusion in systematic reviews [187], [189], [191]. Naive Bayes [192] is another example that works well with bag-of-words features, assuming conditional independence between terms [193], [194]. Logistic regression [195] on the other hand, offers interpretability and adjustable thresholds, making it particularly useful with N-grams or embeddings [196], [197].

The ML and feature extraction techniques described above underpin a range of semi-automated methods designed to streamline systematic review screening by reducing manual effort and improving

efficiency. Software tools like Rayyan [198], Abstrackr [191], RobotAnalyst [199], and others [184], [200], [201], [202], [203] incorporate methods such as N-grams, TF-IDF, and SVM classifiers to transform text into numeric representations, enabling early identification of relevant studies. For instance, Rayyan uses N-grams and SVM to classify studies based on keywords and citations. Abstrackr and RobotAnalyst both utilize TF-IDF for feature extraction, with Abstrackr applying SVM for classification, while RobotAnalyst also combines bag-of-words and LDA. For a more detailed analysis of these tools, along with their underlying methods, see [200].

4.1.3 Machine Learning Tools and AI Applications in Systematic Review Screening: Since its launch in November 2022, OpenAI's ChatGPT has shown potential for use in a number of applications, though its performance varies across domains and may be subject to limitations, as noted in existing literature [139], [149], [150], [151], [152], [153], [156]. Yet, despite these limitations, the performance and reduced cost of AI tools in recent years, especially Large Language Models (LLMs), continue to hold promise for transforming fields such as biomedicine and informatics. Recent advancements in language models, such as GPT-3.5 Turbo, GPT-4, and newer models like 4o-mini and the o1 series, have expanded the potential for data-intensive tasks, analysis, synthesis, and interpretation within systematic review processes [168], [204], [205], [206]. These models represent a promising evolution in ML tools for evidence synthesis, with potential to bridge the science-policy gap, renew public trust, and reduce delays in implementing research findings into practice [17], [170].

Building on recent advancements, the application of LLMs offers a novel approach to systematic review automation [130], [168], [184]. Unlike traditional ML, which requires extensive labeled data and manual feature engineering, LLMs can interpret screening criteria with minimal training data. Our lab's pre-print study showed that GPT-3.5 Turbo performed well in article screening across topics, though sensitivity was a challenge [206]. To address this, we introduced an ordinal prompting strategy, where the model rates its confidence in including each article, allowing for finer adjustments in sensitivity and specificity. This few-shot approach uses minimal contextual information, setting it apart from earlier methods [65]. Additionally, this is the first application of ordinal prompting for review automation, offering a more adaptable tool for data management in systematic reviews. Further details are in the methods and results sections.

4.2 Methods

4.2.1 Overview of Review Data: This dataset comprises 24,534 research articles previously pooled from three systematic reviews across medical and health science domains: Gait and Dementia Signatures (N = 3,245) [159], Cuffless Blood Pressure Monitoring Devices (N = 422) [160], and the Impact of Age and

Comorbidities on Long-COVID (N = 20,867) [161]. These reviews were conducted prior to our introduction to ML tools, with pairs of human reviewers responsible for screening titles, abstracts, and full-text articles. A third reviewer resolved any discrepancies, and we adhered to the guidelines set out by the Cochrane Collaboration and Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) [55], [57].

4.2.2 Article Screening Protocol and Prompt Templates: We utilized OpenAI's GPT-3.5-turbo-0125 model, integrated with the open-source framework LangChain [162] to automate the article screening phase. Table 4.1 compares the layouts of the two prompt templates used as a substitute for manual selection. These templates, along with the study titles and abstracts, were fed into the model, and the responses cleaned for further evaluation.

Table 4.1 Binary and Ordinal Prompts Used to Elicit Model Responses

Binary Prompt	**Ordinal Prompt**
Criteria: {criteria} Should this study be included in the systematic review based on the criteria provided? Answer with True (for inclusion) or False (for exclusion) **Title:** {title} **Abstract:** {abstract} **Inclusion:** [True/False]	**Criteria:** {criteria} On a scale of 1 to 10, where 1 is least likely and 10 is very likely, how suitable is this study for inclusion in the review based on the listed criteria? **Title:** {title} **Abstract:** {abstract} **Inclusion:** [Rating/10]

Note: Please see supplementary files for full prompt templates used.

4.2.3 Few-Shot Evaluation Protocol and Performance Metrics for LLM Screening: We evaluated the model's ability to identify relevant articles using only the titles and abstracts, a technique we refer to as "few-shot" screening. Unlike traditional few-shot learning, which involves training models with a limited number of examples [139], our approach leverages a pre-trained model to make inclusion decisions based on minimal contextual information.

The model's performance was assessed using accuracy, sensitivity, specificity, precision (PPV), Negative Predictive Value (NPV), F1 Score, and Matthews Correlation Coefficient (MCC). To validate these metrics, we used a "pseudo-K-folds" cross-validation approach, a modified K-fold method involving stratified shuffling and partitioning of data into balanced subsets [177]. This approach, similar to a permutation test, was chosen to provide more nuanced control over validation by minimizing bias and variation across data partitions [207], [208]. We then applied bootstrap resampling to calculate standard

deviations for the various metrics. The performance was analyzed using a one-way Analysis of Variance (ANOVA) with an alpha level of 0.05 to determine statistical significance.

4.*2.4 Evaluating Ordinal Prompting with Receiver Operator Characteristic Curves:* In a typical ML binary classification workflow, a model's output probability lies on a continuous scale, and an activation function maps it to one of two classes. Receiver Operating Characteristic (ROC) curves plot the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold settings, helping to assess the effectiveness of different classification cut-offs.

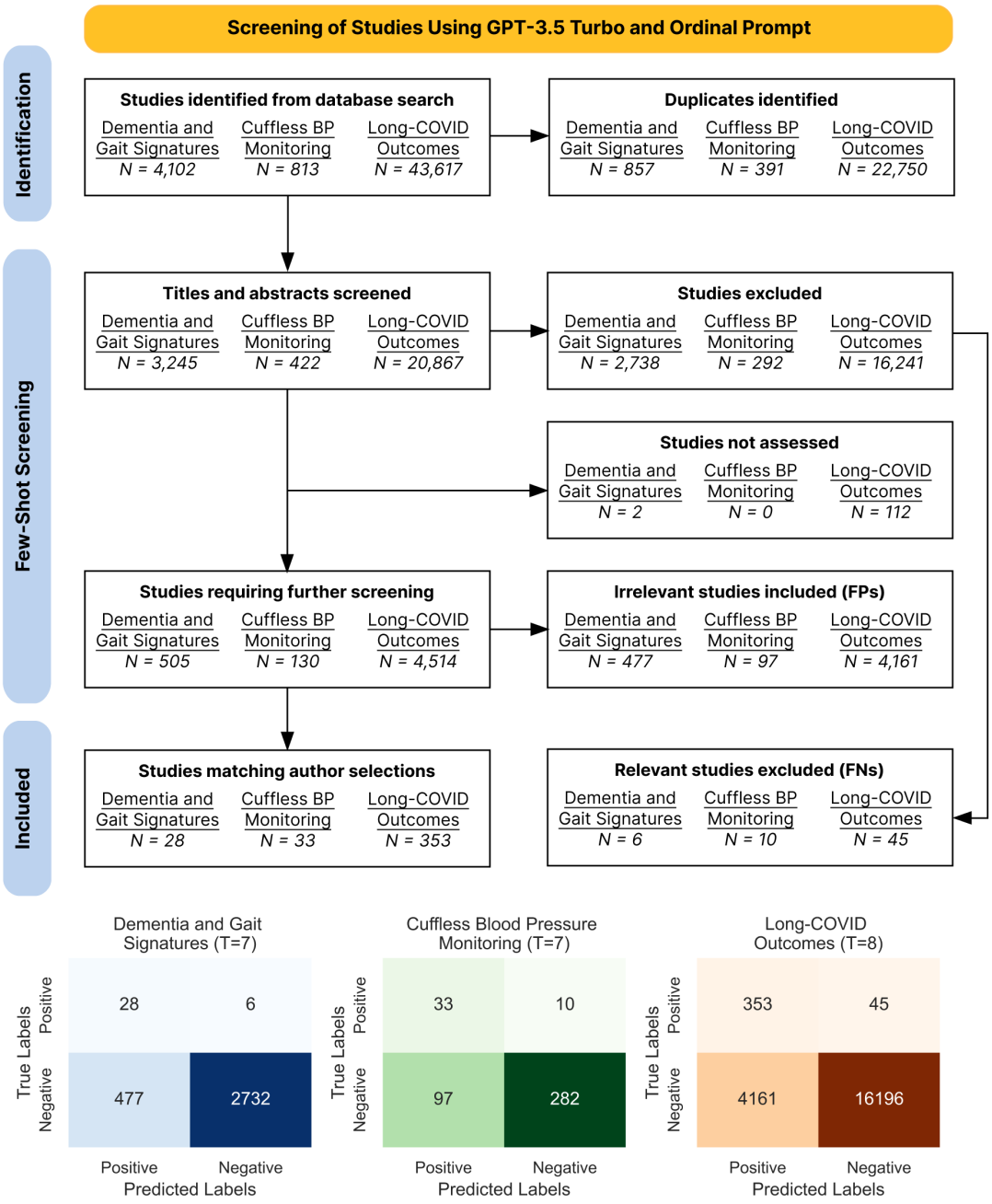
We used ROC curves to evaluate the effectiveness of ordinal prompting by systematically adjusting the threshold from 1 to 10. The Area Under the Curve (AUC) for each review topic was calculated using the Python Scikit-Learn library. The optimal threshold was determined by locating the point on the ROC plot closest to (0,1), indicating the best balance of sensitivity and specificity [209]. Performance metrics for these settings are provided in the supplementary materials as eFigure 2.

4.2.5 Data Transparency and Handling in AI Models: Ethics board approval was not required because the source data was derived from studies with their own ethical approvals.

4.3 Results

4.3.1 Results of LLM-Assisted Study Screening: The PRISMA flow chart in Figure 4.1 shows the steps involved in screening studies using GPT-3.5 Turbo and ordinal prompting. After removing duplicates and conducting few-shot screenings, the process excluded a number of irrelevant studies and minimized false positives and false negatives. The bottom panel displays confusion matrices associated with the screening protocol. This streamlined approach ultimately matched a considerable number of studies with author selections, highlighting the potential of LLMs to reduce manual workload in systematic reviews.

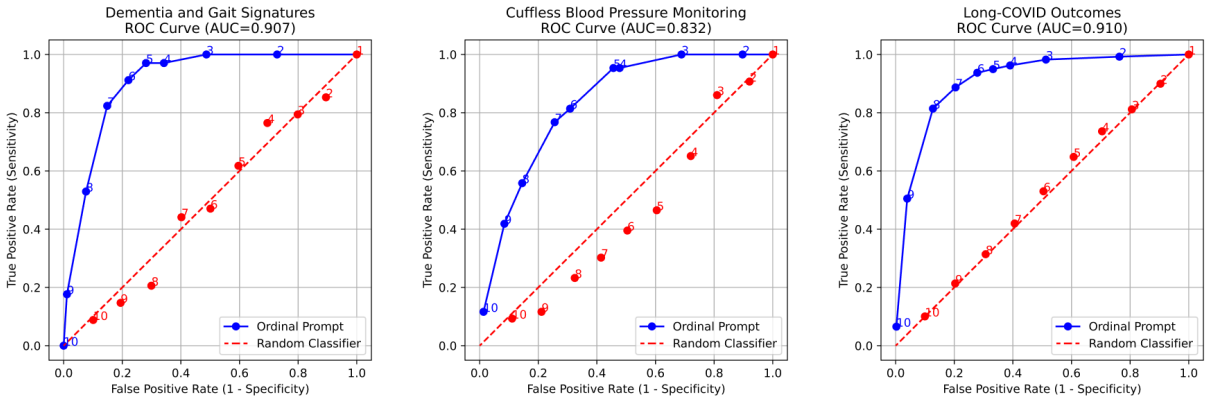
4.3.2 Evaluation of Ordinal Prompting: Figure 4.2 displays ROC curves for various review topics. The AUCs for ordinal prompting ranged from 0.83 to 0.91, demonstrating robust performance across reviews while random classifications did no better than chance. The optimal performance, indicated by the point closest to the top-left [209], consistently ranged between thresholds 7 and 8, and were used as thresholds for the subsequent analysis.



Note: PRISMA flow diagram representing model selections for article screening. Data show results using thresholds T=7, T=7, T=8 respectively. BP: Blood Pressure; FP: False Positive; FN: False Negative; T: Threshold.

Figure 4.1 PRISMA Flow Diagram and Confusion Matrices set at Optimal Threshold for Each Review

In most metrics, no significant improvement was observed over the binary prompt; however, accuracy and specificity were notably greater in one review. It is also worth noting that in the Cuffless BP Monitoring review, the binary prompt achieved a significantly higher specificity than the ordinal prompt, highlighting the mixed results across different scenarios.



Note: Receiver Operator Characteristic (ROC) curves with Area Under the Curve (AUC) representing performance. The red trace represents a single trial from a random multi-class classification control.

Figure 4.2 ROC Curves for Evaluating Optimal Ordinal Threshold Compared to Random Multi-Class Control

4.3.3 Presentation of Optimal Performance Metrics: Table 4.2 presents a comparison of performance metrics for the binary and ordinal prompting strategies. While the ordinal prompting strategy demonstrated a trend toward less variability in sensitivity across topics, the extent of this variability reduction was not uniform and would benefit from quantification in future work.

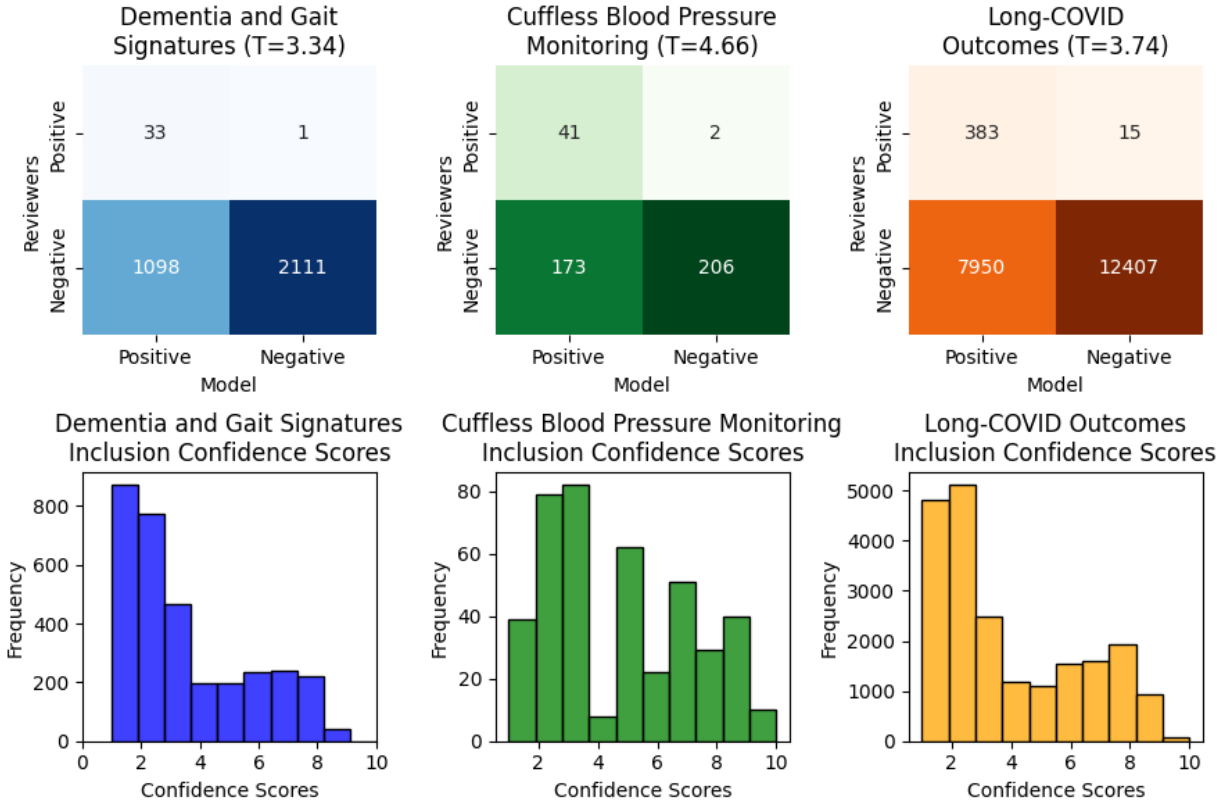
Table 4.2 Performance Metrics Comparing Optimal Ordinal Threshold to Binary Prompt

Metric	Dementia and Gait Signatures (10-fold)		Cuffless BP Monitoring (10-fold)		Long-Covid Outcomes (10-fold)	
	Binary	**Ordinal**	**Binary**	**Ordinal**	**Binary**	**Ordinal**
Accuracy	0.77 ± 0.02	**0.85 ± 0.02**	0.79 ± 0.05	0.75 ± 0.05	0.87 ± 0.01	0.87 ± 0.01
Sensitivity	0.95 ± 0.16	0.87 ± 0.16	0.75 ± 0.21	0.81 ± 0.22	0.86 ± 0.08	0.82 ± 0.07
Specificity	0.76 ± 0.02	**0.85 ± 0.02**	**0.80 ± 0.05**	0.74 ± 0.05	0.87 ± 0.01	0.87 ± 0.01
PPV	0.04 ± 0.02	0.05 ± 0.02	0.29 ± 0.15	0.25 ± 0.09	0.11 ± 0.03	0.11 ± 0.02
NPV	1.00 ± 0.00	1.00 ± 0.00	0.96 ± 0.03	0.97 ± 0.04	1.00 ± 0.00	1.00 ± 0.00
F1 Score	0.08 ± 0.03	0.10 ± 0.03	0.39 ± 0.16	0.37 ± 0.13	0.20 ± 0.05	0.20 ± 0.04
MCC	0.16 ± 0.05	0.19 ± 0.04	0.36 ± 0.14	0.34 ± 0.14	0.28 ± 0.05	0.27 ± 0.04

Note: Metrics were derived using a pseudo-K-folds cross-validation method. BP – Blood Pressure, PPV – Positive Predictive Value, NPV – Negative Predictive Value, MCC – Matthews Correlation Coefficient. Optimal threshold scores were used as cut-off (T=7, T=7, T=8 respectively). Bolded values indicate a p-value < .05.

4.3.4 Analysis of Model Predictions and Confidence Scores: Figure 4.3 presents confusion matrices along with histograms of the model's confidence scores, which tended to be right-skewed. The mean confidence scores, ranging between 3.3 and 4.7, were utilized as the cut-off thresholds for the confusion matrices depicted in the upper panel.

4.3.5 Assessing Few-Shot Performance Using Mean Score: Table 4.3 presents a detailed comparison based on the mean scores for the two different prompting strategies. The results for ordinal prompting were mixed. Although sensitivity consistently showed significant improvement, both accuracy and specificity were lower, suggesting a steep trade-off between sensitivity and specificity.



Note: Upper panel includes confusion matrices with T as the average confidence score. Bottom panel includes histograms of model decisions.

Figure 4.3 Confusion Matrices set at Average Threshold and Histogram of Confidence Scores

4.3.6 Quantifying Screening Workload Reduction: To quantify the screening workload reduction while ensuring relevant studies were not excessively excluded, we applied the mean threshold and maximized sensitivity. For each review topic, the workload reduction, W, was calculated using the following formula:

$$W = 100\% \times \left(1 - \frac{Included\ Studies\ After\ Filtering}{Total\ Initial\ Studies}\right)$$

Substituting the relevant values from Figure 3, our method yielded workload reductions of approximately 65%, 49%, and 60%, while maintaining sensitivity scores above 93%. Alternatively, applying a higher threshold like those shared in Figure 4.1 could increase workload savings further to approximately 84%, 69%, and 78%, though this would come at the expense of lower sensitivity.

Table 4.3 Performance Metrics of Average Threshold for Ordinal Compared to Binary Prompt

Metric	Dementia and Gait Signatures (10-fold)		Cuffless BP Monitoring (10-fold)		Long-Covid Outcomes (10-fold)	
	Binary	**Ordinal**	**Binary**	**Ordinal**	**Binary**	**Ordinal**
Accuracy	**0.76 ± 0.02**	0.66 ± 0.01	**0.79 ± 0.05**	0.58 ± 0.07	**0.87 ± 0.01**	0.62 ± 0.01
Sensitivity	0.95 ± 0.16	0.98 ± 0.05	0.75 ± 0.21	**0.93 ± 0.16**	0.86 ± 0.08	**0.96 ± 0.03**
Specificity	**0.76 ± 0.02**	0.66 ± 0.01	**0.70 ± 0.05**	0.54 ± 0.06	**0.87 ± 0.01**	0.61 ± 0.01
PPV	0.04 ± 0.02	0.03 ± 0.02	0.29 ± 0.15	0.19 ± 0.09	**0.11 ± 0.03**	0.05 ± 0.01
NPV	1.00 ± 0.00	1.00 ± 0.00	0.96 ± 0.03	**0.99 ± 0.02**	1.00 ± 0.00	**1.00 ± 0.00**
F1 Score	0.08 ± 0.03	0.06 ± 0.03	0.39 ± 0.16	0.31 ± 0.12	**0.20 ± 0.05**	0.08 ± 0.01
MCC	0.16 ± 0.05	0.13 ± 0.03	0.36 ± 0.14	0.29 ± 0.12	**0.28 ± 0.05**	0.16 ± 0.01

Note: Metrics were derived using a pseudo-K-folds cross-validation method. BP – Blood Pressure, PPV – Positive Predictive Value, NPV – Negative Predictive Value, MCC – Matthews Correlation Coefficient. Mean threshold scores were used as cut-off (T=3.34, T=4.66, and T=3.74 respectively). Bolded values indicate a p-value < .05.

4.3.7 Summary of Model Inclusion Scoring for Long-COVID Studies: Table 4.4 provides inclusion scores, *S*, assigned by GPT-3.5 Turbo, with the review on long-COVID health outcomes used to illustrate our approach. Each score reflects how well a study meets eligibility criteria, including focus, design, and relevance to long-term COVID effects, with greater scores being more relevant. The rationale explains the model's assessment for each score, supporting a clearer selection process for the review.

4.4 Discussion

4.4.1 Overview of Findings: This study builds on prior work [206] exploring LLMs in systematic review automation, where initial findings showed both potential and limitations, with error rates around 35% and some relevant studies excluded. Here, we introduce a refined prompting strategy that enhances the model’s capacity to prioritize relevant articles, addressing prior limitations by improving filtering accuracy and usability in systematic review contexts. This advancement supports efficient and dependable literature screening, potentially expediting systematic review workflows, increasing reliability, and aiding in the timely integration of research findings into practice.

Table 4.4 Rationale for Inclusion Scores Assigned by GPT-3.5 Turbo in Systematic Review Screening of Long-COVID Health Outcomes

S	Ordinal Prompt	GPT-3.5 Turbo Reasoning
2	***Title:*** Late onset infectious complications and safety of tocilizumab in the management of COVID-19. [210] ***Abstract:*** BACKGROUND: Tocilizumab (TCZ) has been used in the management of COVID-19-related cytokine release syndrome (CRS). Concerns exist regarding the risk of infections and drug-related toxicities. We sought...	This study does not meet the eligibility criteria for inclusion in a systematic review focused on long COVID health outcomes. The study primarily focuses on the safety and complications of tocilizumab in the management of COVID-19, rather than reporting on health effects or symptoms occurring at least four weeks after a COVID-19 diagnosis. Additionally, the study design does not align with the acceptable study designs for inclusion in the systematic review.
4	***Title:*** Histopathologic evaluation of placentas after diagnosis of maternal severe acute respiratory syndrome coronavirus 2 infection. [211] ***Abstract:*** BACKGROUND: The impact of maternal severe acute respiratory syndrome coronavirus 2 infection on placental histopathology is not well known. OBJECTIVE: To determine if any significant placental...	While the study focuses on the impact of maternal SARS-CoV-2 infection on placental histopathology, it does not specifically address long COVID health outcomes as required for inclusion in a systematic review focused on this topic. The study design is a retrospective cohort study, which is an acceptable design for inclusion. However, the study only looks at placentas from patients diagnosed with SARS-CoV-2 infection during the third trimester of pregnancy and does not report on health effects or symptoms occurring at least four weeks after a COVID-19 diagnosis. Therefore, this study may not be suitable for inclusion in a systematic review focused on long COVID health outcomes.
6	***Title:*** Follow-Up of Coronavirus Infected Patients Using Telemedicine in a Referral Pulmonary Center. [212] ***Abstract:*** BACKGROUND: Telemedicine is considered an innovative approach for management and follow up of communicable diseases, when person to person contact has the risk of disease dissemination, such as the...	While this study focuses on follow-up of coronavirus-infected patients using telemedicine, it does not explicitly mention long-term health effects or symptoms occurring at least four weeks after COVID-19 diagnosis. The study design is not specified, and there is limited information on the specific health outcomes being measured. Additionally, the abstract does not indicate whether the study is peer-reviewed. However, the use of telemedicine for patient follow-up in the context of a communicable disease like COVID-19 could potentially provide valuable insights into long COVID health outcomes. More details from the full study would be needed to make a final determination on its eligibility for inclusion in a systematic review on long COVID health outcomes.
8	***Title:*** Echocardiographic Characteristics of Subjects With COVID-19: A Case Series. [213] ***Abstract:*** Although coronavirus disease 2019 (COVID-19) manifests in most cases with respiratory symptoms, other presentations can occur. Direct damage to the cardiovascular system has been reported and recently...	This study on echocardiographic characteristics of subjects with COVID-19 seems suitable for inclusion in a systematic review focused on long COVID health outcomes. The study reports on cardiovascular manifestations of COVID-19, which may be relevant to long-term health effects. It is a peer-reviewed original research case series, providing valuable data on echocardiographic findings in patients with COVID-19. The study design is appropriate for the scope of the systematic review. The study utilizes echocardiography as a method of symptom measurement, which is acceptable. Overall, this study appears to be a strong candidate for inclusion in a systematic review on long COVID health outcomes.
10	***Title:*** Post-discharge health status and symptoms in patients with severe COVID-19 [214] ***Abstract:*** BACKGROUND: Little is known about long-term recovery from severe COVID-19 disease. Here, we characterize overall health, physical health and mental health of patients one month after discharge for severe...	This study meets the criteria for inclusion in a systematic review on long COVID health outcomes. It reports on health effects occurring at least four weeks after a COVID-19 diagnosis, is peer-reviewed original research, and utilizes longitudinal data collection methods through a cohort study design. The study provides relevant data on post-discharge health status and symptoms in patients with severe COVID-19, including physical and mental health outcomes. The data collected through validated survey instruments meets the criteria for acceptable methods of symptom measurement. Overall, this study is highly suitable for inclusion in a systematic review on long COVID health outcomes.

Note: *S* denotes the inclusion suitability score assigned by GPT-3.5 Turbo based on study relevance to long-COVID health outcomes; the full prompt template is provided in the supplementary materials.

4.4.3 Addressing Resource Allocation and Cost-Efficiency: Systematic reviews traditionally require considerable time and resources, which can lead to redundancy in certain research areas while other topics remain under-researched. Our system, costing approximately $0.31 per 1000 studies screened, demonstrates an economical approach that supports sustainable replication at scale, addressing wasteful resource use. This method empowers reviewers to maintain high standards with lower costs, enhancing the accessibility and reliability of systematic reviews. By reducing the dependency on extensive computational resources and specialized training, we enable researchers to fact-check critical review areas more readily, reducing inefficiencies without compromising rigor.

4.4.4 Adjustability and Context-Specific Application in Research: Given the breadth of topics in medical and scientific literature, systematic review tasks demand adaptable screening tools that can be tailored to diverse sensitivity and specificity requirements. A key advantage of our approach lies in its adjustability, allowing users to balance inclusion sensitivity with specificity according to the review context. For instance, in infectious disease monitoring, prioritizing higher sensitivity ensures potential cases are not overlooked, while in systematic reviews focused on niche medical treatments, high specificity can help exclude irrelevant studies. By leveraging an ordinal prompting strategy, our model fine-tunes responses for relevance, aligning well with real-world needs where context-specific flexibility is needed.

4.4.5 Strengths and Comparison with Existing Screening Tools: Traditional ML approaches, while effective, often rely on large, annotated datasets, which limits their scalability across varied topics. Moreover, ML screening algorithms with heuristic stopping criteria reduce workload but depend on supervised data, which limits flexibility in adapting to new research questions [215]. For example, Rayyan, a widely used semi-automated screening tool, applies fixed thresholds for inclusion/exclusion decisions. While such thresholds help standardize decisions, they may reduce precision in complex reviews where nuanced interpretation is required [216]. Similarly, Abstrackr achieves high accuracy with extensive training on established topics, but its reliance on annotated data restricts adaptability for emerging areas [167], [217]. RobotReviewer, used in clinical trial bias assessment, requires manually labeled data for initial training, reducing its generalizability outside clinical contexts [203].

In contrast, LLMs using few-shot or ordinal prompting can adapt more easily to evolving criteria or subtle differences in study relevance. However, this flexibility introduces its own risks, particularly the possibility of subjective bias in prompt engineering or inconsistent behavior across different topics. As such, while our results demonstrate the potential for more context-sensitive interpretation, we caution that LLM outputs should be treated as decision support tools, not replacements for reviewer judgment. Acknowledging these trade-offs is essential to responsibly integrating AI into evidence synthesis.

4.4.6 Weaknesses, Constraints and Methodological Considerations: Our approach, while adaptive, has limitations. Given the rapid release of new models and frequent changes to meta-parameters such as temperature and other settings, performance outcomes are not readily generalizable across contexts. The LLMs discussed here are largely black-box and closed-source models, presenting challenges for thorough evaluation and interpretability by researchers. Additionally, our use of bootstrap sampling and permutation testing though helpful for estimating variability, may elevate the risk of Type I errors (false positives) and may not fully capture population variance [207], [208].

To support generalizability, we validated our protocol across three distinct review topics, encompassing a total of 24,534 articles. We used a 10-fold partitioning scheme during validation for better power analysis, testing different fold counts internally. Finally, we interpreted performance in conjunction with ground truth reviewer data and qualitative inspection of false positives and negatives, reducing the likelihood of over-interpreting noise as signal. Despite these safeguards, some caution is still warranted when interpreting results across different datasets and model contexts. Future studies may benefit from using ensemble models across different LLM providers (e.g., OpenAI, Meta, Google, DeepSeek) and varied prompt designs to better assess consistency and reduce the risk of model-specific bias.

4.4.7 Future Research Directions: Emerging models such as GPT-4o and the o1 series offer new directions for improving cost-efficiency and reasoning capabilities in review workflows. GPT-4o, released in mid-2024, is approximately 60% cheaper than GPT-3.5 Turbo [218], making large-scale deployment more accessible. Meanwhile, the o1 series, currently in preview, adapts computational load based on task complexity, potentially making it better suited for high-variance research tasks [219]. Although these models are still being rolled out, their evolution signals the growing feasibility of end-to-end AI-assisted systematic reviews.

Current efforts are focused on evaluating LLMs for full-text data extraction. Future research should explore integrating AI across all phases of systematic reviews, extending beyond initial screening to include end-to-end analysis. Complementary methods like the utilization of grey literature [220] or publication type tagging [221] could also contribute to the improvement of review automation. Future steps to better identify and synthesize information, particularly from the former could potentially transform the systematic review process by incorporating a broader range of data.

4.5 Conclusion

This study demonstrates the potential of LLMs, particularly GPT-3.5 Turbo, to enhance systematic review processes by using scalable prompting strategies that balance sensitivity and workload reduction. Our application of binary and ordinal prompts achieved substantial screening efficiency gains, reducing manual workload by 49-65% across three large systematic reviews without compromising sensitivity. Additionally, the innovative use of ordinal prompts allowed for nuanced control over study inclusion, facilitating greater adaptability in diverse review contexts. This adaptability, combined with the capacity of LLMs to handle large-scale data with minimal training, positions LLMs as valuable tools for accelerating evidence synthesis in systematic reviews. Finally, we highlight a practical pathway for AI integration that not only reduces resource strain but also improves transparency and replicability in evidence-based research.

5. Algorithms, Architecture, and Design Patterns for Smartphone-Based Mobility Tracking

The previous two chapters demonstrated how large language models can streamline and enhance evidence synthesis, providing a scalable approach to managing the extensive corpus of biomedical literature. With these advancements, we now shift focus to tangible, everyday technologies that translate personalized health monitoring from theory to practice.

Chapter 5 turns to smartphones—ubiquitous, sensor-rich devices that hold immense potential in healthcare monitoring. If Chapter 2 highlighted gait analysis as a diagnostic tool and Chapters 3 and 4 focused on improving the evidence landscape, this chapter grounds our discussion in the reality of implementation. Specifically, it details the design and development of a smartphone application for collecting and managing data from built-in sensors.

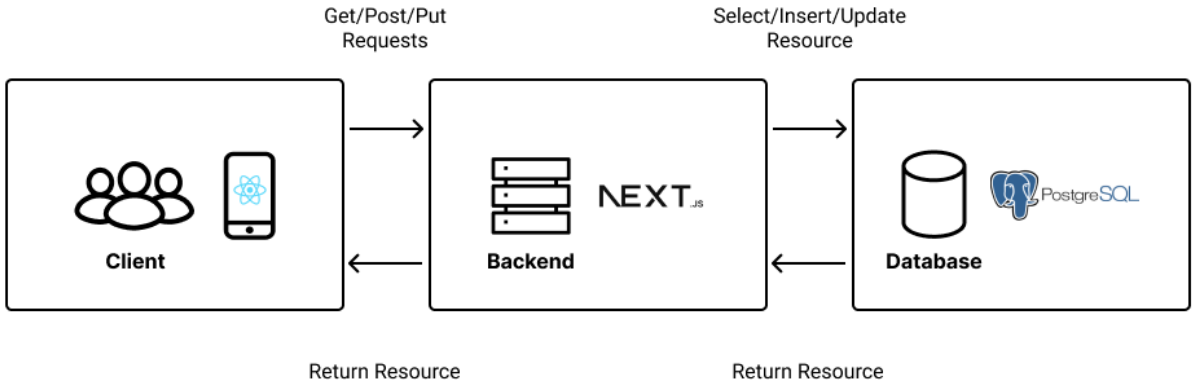
In doing so, we lay the groundwork for smartphone-based human activity recognition (HAR) as a solution that can bridge the gap between research findings and patient-level health insights. While the system was developed with older adult populations in mind, early testing was conducted using a convenience sample of younger participants. As such, this chapter emphasizes system feasibility and design rather than clinical validation. Future studies will be needed to evaluate usability and model performance in aging populations under real-world conditions.

5.1 System Overview

5.1.1 System Design and Architecture Choices: Our application architecture consisted of three main components: the client device, the backend server, and the database. We opted for a React Native [222] client using Expo [223] for frontend development to allow for cross-platform compatibility between iOS and Android. The backend API endpoints followed RESTful design principles and were implemented using the Next.js framework [224]. We also used a(n) SQLite database for the data storage schema and models, with Prisma ORM [225] serving as the object-relational mapping tool. During deployment, we migrated the database to PostgreSQL, which was better suited for cloud and edge computing environments. Finally, the backend server was hosted on Vercel [224], chosen particularly for its Next.js ecosystem support. Our software architecture is illustrated in Figure 5.1 below.

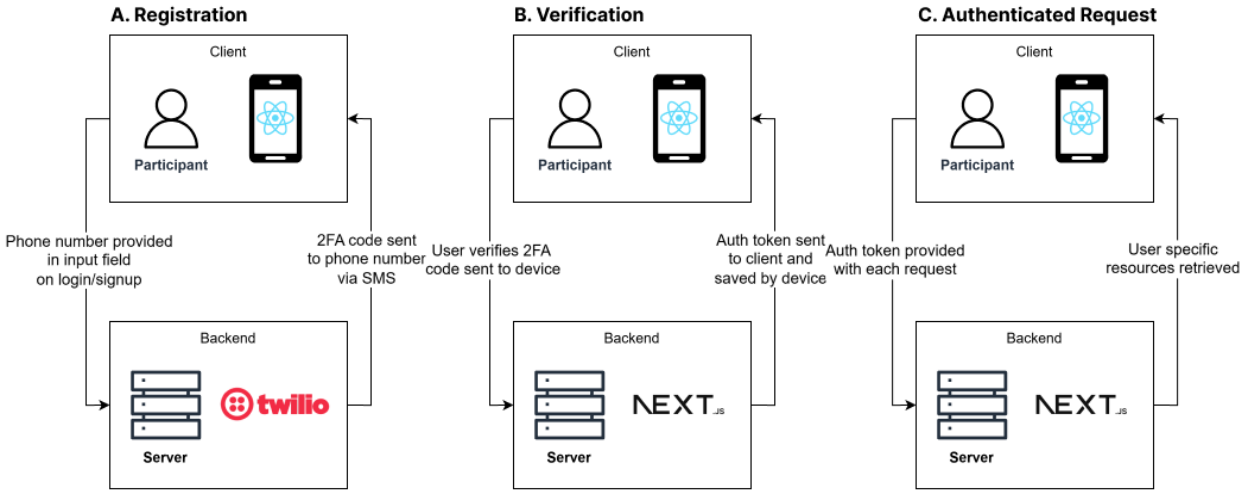
5.1.2 User Authentication and Security Concerns: Our security architecture was designed with three primary goals: verifying the identity of real individuals, eliminating the risks of password storage, and

enabling secure communication with study participants. To achieve these objectives, we implemented a cellular two-factor authentication (2FA) system, as illustrated in Figure 5.2.



Note: Communication between the client and the backend occurs via HTTP methods such as GET, POST, and PUT requests, with the backend processing these requests and interacting with the PostgreSQL database for data persistence.

Figure 5.1 Architecture Component Diagram



Note: This diagram illustrates the three-step user authentication process: (A) Registration using Twilio to send a 2FA code via SMS, (B) Verification where the user enters the code to receive an encrypted authentication token, and (C) Authenticated Requests where the token validates user identity and retrieves user-specific resources securely.

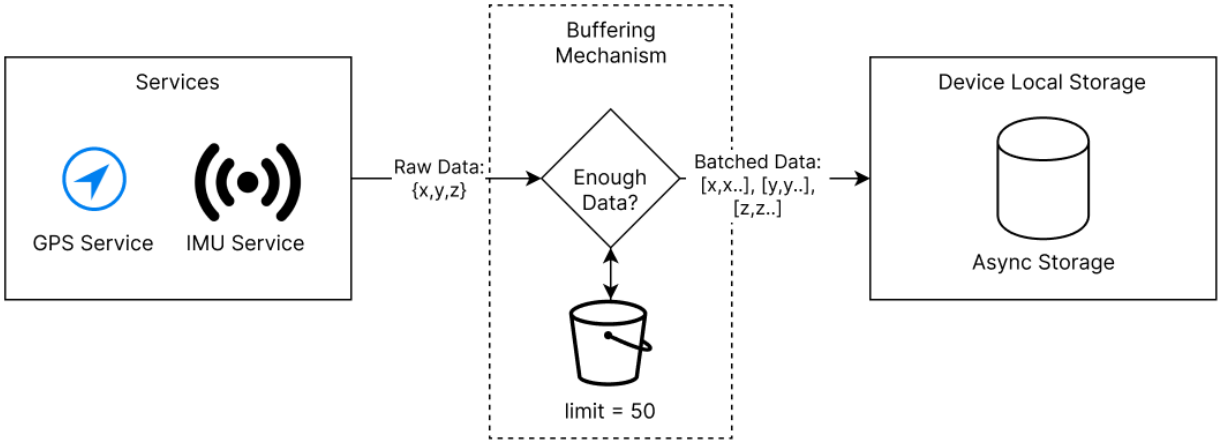
Figure 5.2 Registration, Verification, and Authentication Flow

The process begins with registration (Figure 5.2A). For this purpose, Twilio sends a randomly generated four-digit code via SMS to the user's phone number at a cost of just $0.0079 per message. Upon receiving the code, the user verifies device ownership (Figure 5.2B) by entering the code into the

application. Then, the backend server issues a personalized, encrypted authentication token, which the device stores and uses to authenticate the user's identity in future requests. These tokens contain identity information but are encrypted with a server-privileged key. During subsequent interactions, the backend server acts as an intermediary to the database, using the token to ensure that queries retrieve data specific to the authenticated user (Figure 5.2C).

5.1.3 Data Collection and Sensor Support: Client data collection and sensor support were divided into two distinct services. The first service managed the sampling of IMU sensors, specifically the three-axis accelerometer, gyroscope, and magnetometer. The second service handled data collection from the GPS module.

To manage the high-frequency data generated by these services, we implemented a buffering mechanism. This approach avoided the inefficiencies of frequent read/write operations to the device's storage, which would occur at sampling rates of roughly 20 to 50Hz. Instead, incoming data points were temporarily stored in a buffer. Once the buffer reached a predefined threshold, the data was written to storage in batches (Figure 5.3).



Note: This diagram illustrates the batching mechanism used to save sensor data to local device storage. GPS: Global Positioning System; IMU: Inertial Measurement Unit; Async Storage: Asynchronous Device Storage.

Figure 5.3 Client Buffering and Storage Infrastructure

Our design logic extended to the API endpoints and data schema. For example, each activity was represented as a SensorSession, which encapsulated SensorData and LocationData. These components were managed through dedicated API endpoints, enabling independent uploads of SensorData and LocationData while maintaining their association within the same session.

This design also accounted for two critical requirements: the potential loss of internet connectivity and the need to handle large and frequent data transfers. In the case of a stable internet connection, the system would allow for pseudo-streaming by sending batched SensorData and LocationData packets to the server. In the case of unstable or no internet connection, data could be stored locally and synced with the server upon a re-established connection.

5.1.4 Database Schema and Back-of-the-Envelope Calculations: The database schema (Figure 5.4) contained five interrelated tables: Participant, SensorSession, SensorData, HealthMetrics, and LocationData. The SensorData table, storing high-frequency readings, was the largest contributor to storage requirements, while other tables provided essential metadata, participant details, and contextual information for activity tracking.

To estimate the data storage requirements for the system, we performed simple calculations based on the schema in Figure 5.4. Below, we focus on SensorData since it constituted the largest storage requirements.

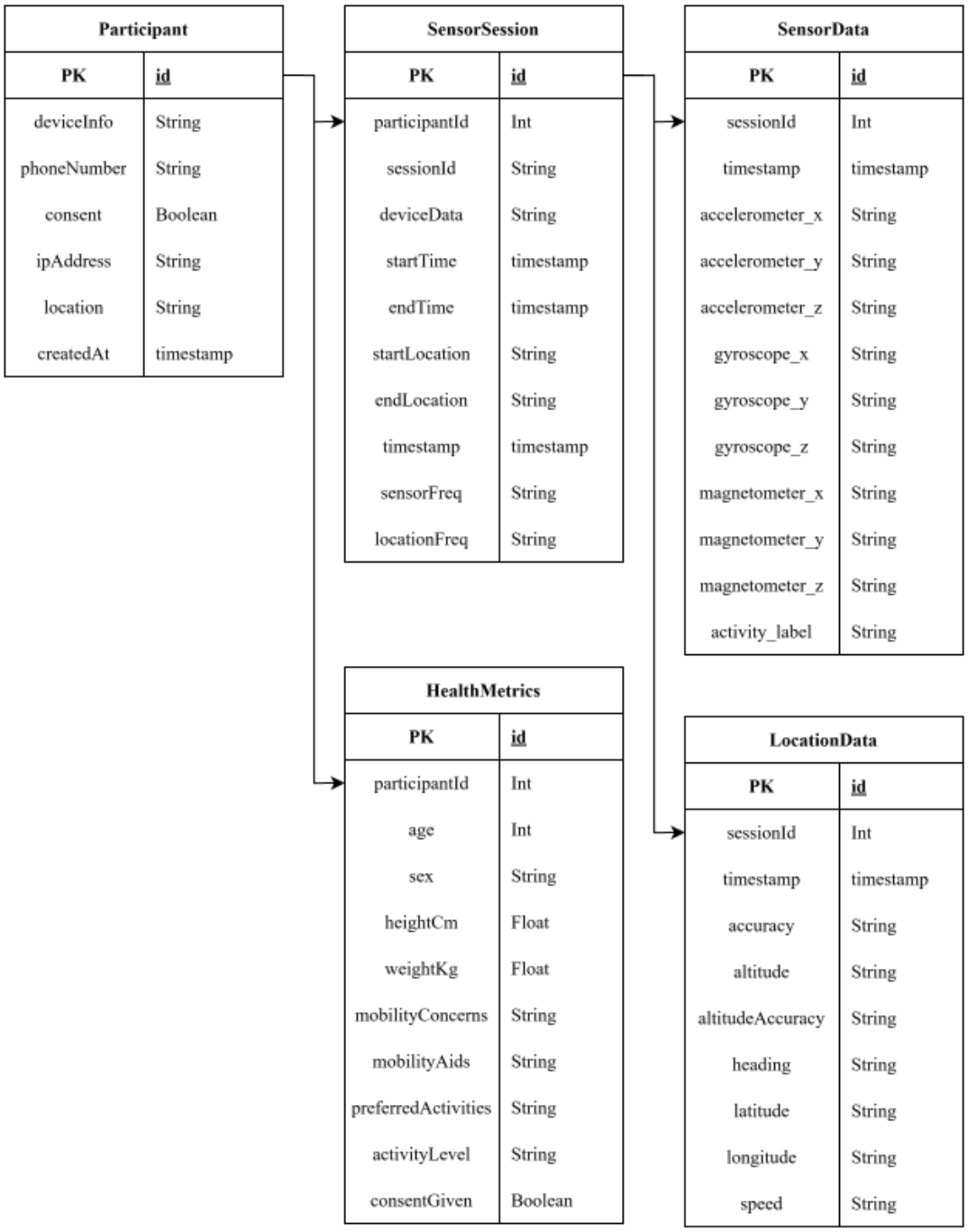
Assuming a sampling rate of 50Hz, each second of data collected from the three-axis IMU sensors (accelerometer, gyroscope, magnetometer) would include 450 floating-point values. Assuming each value to be about 4 bytes, each second would contain 1.8 KB of sensor data. Over a typical 20-minute activity, this amounts to approximately 2.16 MB of sensor data, with an additional 116 bytes for metadata fields such as timestamps, activity labels, and session identifiers. For a single user recording one session per day, the total storage requirement is approximately 2.16 MB per day.

Extrapolating to 100 users over 30 days, the total storage requirement was estimated to be approximately 6.48 GB. Extending this to a year, the requirement scales to approximately 78.84 GB. Additionally, given Vercel's pricing model of $0.12 per GB for read and write operations, the cost of supporting 100 users for 30 days was estimated to be $1.56. For a full year, this would amount to approximately $18.92 in storage costs. These estimates demonstrate the cost-effectiveness of supporting small to medium-scale studies and research projects.

5.2 Software Validation Methodologies

5.2.1 Device and Environment Testing: To ensure compatibility across both Android and iOS platforms, we tested the application on two smartphone devices: a Samsung Galaxy S10e (2019) and an iPhone 14 Pro (2022). Testing on devices released three years apart provided confidence that our software would function effectively on both newer and older hardware. Additionally, we compared the results to a third-party IMU device [226] (MetaMotionS; MbientLab) to confirm the accuracy of the smartphone sensors

against a reliable benchmark. However, due to differences in configurable sampling frequencies, comparisons with the IMU device were limited to motion-related tests.



Note: This diagram shows the database schema, where Primary Keys (PK) uniquely identify records across tables such as Participant, SensorSession, SensorData, HealthMetrics, and LocationData.

Figure 5.4 Entity Relationship Diagram of Postgres Database

Our testing methodology followed a tiered deployment strategy, in which different environments were used for iterative testing and validation. In the context of our application, "staging" referred to a pre-production environment where new features were deployed and tested before being released to production. This setup allows for thorough evaluation without affecting the live environment. However, as the application is still under review and a production environment does not yet exist, all tests described in this chapter were performed in the staging environment. Furthermore, while the application collects both IMU and GPS data, this chapter primarily focuses on IMU data from the smartphones. GPS data was excluded from analysis due to privacy concerns.

5.2.2 Data Sampling Validation: Sensor readings are processed as asynchronous calls, meaning they may not always align precisely with the desired sampling frequency due to variations in hardware, software, and background processes. To validate the actual sampling rate against the set rate, both devices were laid flat on a surface and set to record for 30 seconds. The sampling frequency in the application was controlled by specifying the time elapsed between sensor readings. For instance, a sampling rate of 20Hz corresponds to a sensor interval of 50ms, calculated using the formula:

$$f = \frac{1000ms}{\Delta s}$$

where f is the desired sampling frequency in Hz, and Δs is the sampling interval in milliseconds. After recording, the actual sampling frequency was determined using the equation:

$$f_{actual} = \frac{n}{t}$$

where n represents the total number of samples recorded, and t is the total elapsed time of the session in seconds. This stationary test was repeated for five trials at different sampling frequencies and analyzed using a two-sample T-test with unequal variance to identify any significant discrepancies between the devices.

5.2.3 Drop Test Methodology: Accelerometers on the devices measure acceleration in terms of g values, where 1g (9.81 m/s²) represents the gravitational force exerted by Earth's gravity. The true acceleration (a_{actual}) can be calculated using the equation:

$$a_{actual} = g \cdot s$$

where g is the gravitational acceleration (9.81 m/s²), and s is the scale factor reported by the accelerometer. To supplement our stationary test, we conducted a motion test by dropping both devices

from a height of ~2 meters onto a soft surface. Free fall provided a consistent and reliable test condition for repeated motion experiments, as we know gravitational acceleration to be approximately 9.81 m/s².

Finally, while we tested the devices at sampling frequencies between 20 and 100Hz, no significant differences were observed across frequencies; therefore, our results are primarily reported at 50Hz. To ensure consistency, the devices were oriented in the same direction throughout the test to avoid variations due to external magnetic fields.

5.2.4 Motion and Stability Testing: Building on the stationary and drop tests, we extended our methodology to include real-life examples of gait to simulate everyday activity. To achieve this, both devices were placed in the right pant pocket and set to record during an ~8-meter walk. Each walking trial lasted 15 to 20 seconds and included deliberate pauses at the start and end of the activity to distinguish the walking motion from noise artifacts.

The x, y, and z components of each sensor were combined to calculate the resultant vector magnitude, which provides a scalar representation of overall motion. The resultant was computed using the equation:

$$R = \sqrt{x^2 + y^2 + z^2}$$

where x, y, and z are the acceleration, angular velocity, or magnetic field strength components measured along the three axes. For qualitative analysis, a representative example of the recorded data was selected to highlight key observations and validate the system's ability to capture meaningful walking patterns.

5.3 Performance Metrics

5.3.1 Sampling Frequency Validation: The data revealed that the newer-generation iOS device generally achieved higher and more consistent sampling rates compared to the older Android device, which struggled at higher input frequencies (Table 5.1).

Interestingly, the iOS device exhibited notable variability at specific frequencies, particularly 40Hz and 80Hz. For instance, when the input sampling frequency was set to 80Hz, the accelerometer and gyroscope achieved rates as low as 49Hz in some trials and up to 79Hz in others. While this may be due to intermittent factors or system-level differences, the large variation highlights the potential for occasional inconsistencies even in newer devices.

5.3.2 Drop Test Analysis: The drop test among the test devices showed similar results for the accelerometer (Figure 5.5). For all devices, a resultant acceleration of approximately 1g was recorded at the start of the experiment. At different time points during the trials, the resultant acceleration dropped to 0g, indicating free fall. This was followed by sharp spikes upon impact, caused by the sudden deceleration of the devices hitting the surface. Interestingly, the gyroscope data revealed an upside-down U-shaped pattern during free fall, followed by a sharp spike upon impact, capturing the rotational force experienced during the collision.

Table 5.1 Sampling Rate Discrepancies for Smartphone IMU Sensors at Different Input Frequencies

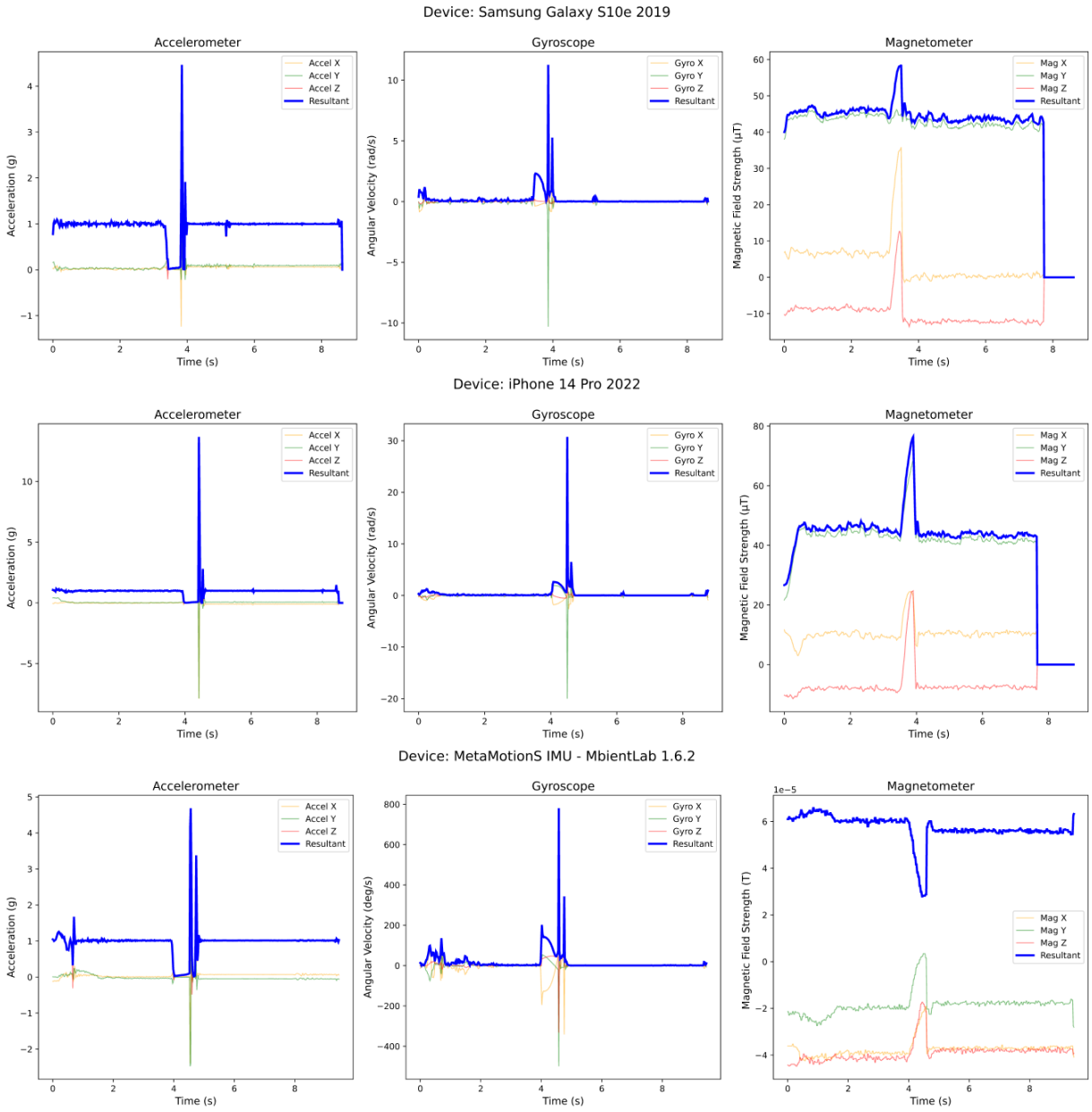
Input Freq (Hz)	Accelerometer Freq (Hz)		Gyroscope Freq (Hz)		Magnetometer Freq (Hz)	
	Samsung Galaxy S10e 2019	**Apple iPhone 14 Pro 2022**	**Samsung Galaxy S10e 2019**	**Apple iPhone 14 Pro 2022**	**Samsung Galaxy S10e 2019**	**Apple iPhone 14 Pro 2022**
20	18.58 ± 0.08	**19.83 ± 0.02**	18.6 ± 0.07	**19.82 ± 0.01**	17.31 ± 0.07	**19.82 ± 0.02**
40	32.13 ± 0.24	33.68 ± 8.11	32.11 ± 0.28	33.69 ± 8.12	30.32 ± 0.24	**39.49 ± 0.19**
60	46.82 ± 0.51	**60.79 ± 4.90**	46.87 ± 0.57	**63.21 ± 2.67**	40.63 ± 0.22	**65.87 ± 0.23**
80	55.84 ± 0.31	61.37 ± 16.18	55.98 ± 0.34	61.36 ± 16.19	45.66 ± 0.26	**49.49 ± 0.09**
100	59.14 ± 0.32	**98.30 ± 0.21**	59.30 ± 0.25	**98.29 ± 0.20**	49.28 ± 0.42	**98.27 ± 0.19**

Note: Cells display mean ± standard deviation of sampling rates of five trials across accelerometer, gyroscope, and magnetometer sensors on the Samsung Galaxy S10e (2019) and iPhone 14 Pro (2022). Bolded values indicate significant differences (p-value < 0.05) from a Student's t-test.

In contrast, the magnetometer data returned to baseline immediately after impact, showing no significant spike. During free fall, all three devices displayed a U-shaped pattern in the magnetometer readings, though the IMU device exhibited an inverted U-shape compared to the smartphones, likely due to differences in sensor calibration or sensitivity. Additionally, the third-party device measured gyroscope data in degrees per second (deg/s), while the smartphone sensors reported angular velocity in radians per second (rad/s). For the magnetometer, the third-party device provided readings in a scaled format (e.g., teslas with scientific notation), whereas the smartphones reported values directly in microteslas.

5.3.3 Motion and Stability Testing: Illustrated in Figure 5.6, the test devices showed cyclical traces during walking trials, reflecting their ability to capture natural gait patterns. The accelerometer and gyroscope readings demonstrated similar magnitudes of the resultant across all devices, indicating consistent performance in capturing dynamic motion across platforms. However, notable differences were observed in the magnetometer readings among the devices. One device recorded resultant magnitudes

nearly double those of the others, which may stem from differences in hardware calibration, sensitivity, or the influence of external magnetic fields on the devices.



Note: Raw accelerometer, gyroscope, and magnetometer readings during free fall drop tests on the Samsung Galaxy S10e (2019), iPhone 14 Pro (2022), and MetaMotionS IMU. The drop event is marked by the near-zero resultant magnitude during free fall, followed by sharp spikes at impact.

Figure 5.5 Sensor Responses During Free Fall and Impact Across Test Devices

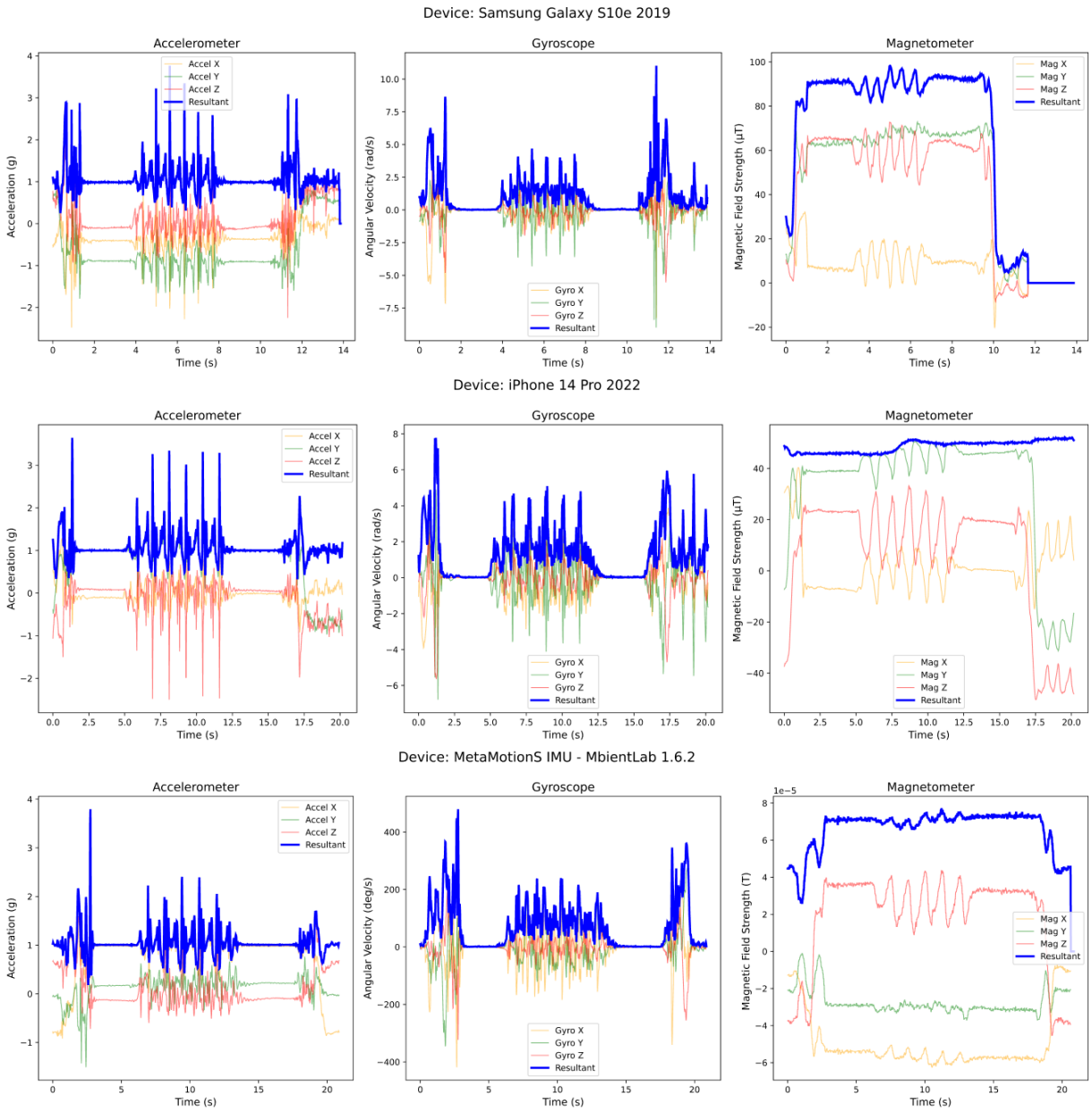
5.4 Challenges and Solutions

5.4.1 Managing High Frequency Sensor Data: The sensors on each smartphone exhibited inherent limitations, as demonstrated in our stationary sampling frequency experiment. For instance, while both devices were set to sample at 100 Hz, the Android device consistently maxed out at around 60 Hz, while the newer iOS device achieved up to 98 Hz. Variability too, was observed both between trials, with differences of up to 30 Hz recorded at the same settings. Between sensors, the magnetometer also displayed lower variability across trials compared to the accelerometer and gyroscope. These inconsistencies were likely influenced by a combination of hardware capabilities, sensor throttling, and software limitations.

One contributing factor may be the asynchronous nature of sensor sampling, where the specified interval represents the minimum time between samples, leading to actual frequencies that are often lower than expected. Additionally, to mitigate battery drain and prolong device usability, operating systems may intentionally throttle high-frequency sensor sampling, particularly for power-intensive sensors.

5.4.2 Device and Platform Heterogeneity: Another challenge involved managing differences between operating systems. React Native and Expo simplified development by allowing a single codebase to be used for both iOS and Android. However, this approach introduced new challenges when features were implemented differently or were available on one platform but not the other. A notable example was how native pedometer modules were implemented. On iOS, we could input two date fields to retrieve the number of steps taken between those times. In contrast, the Android pedometer module functioned more like a live sensor, providing step data only while actively subscribed and recording.

5.4.3 User Interface and Flow: The application was designed for use across a wide range of ages and health contexts, and therefore prioritized accessibility and usability. While primarily intended as a research and IMU data collection tool, the user interface and experience (UI/UX) drew inspiration from fitness and running applications to provide additional utility beyond data collection (Figure 5.7). The inclusion of GPS alongside IMU data enhances its potential for sophisticated analyses, such as correlating walking quality with specific locations. This could enable features like heatmaps to highlight areas that may require accessibility improvements or urban planning updates.

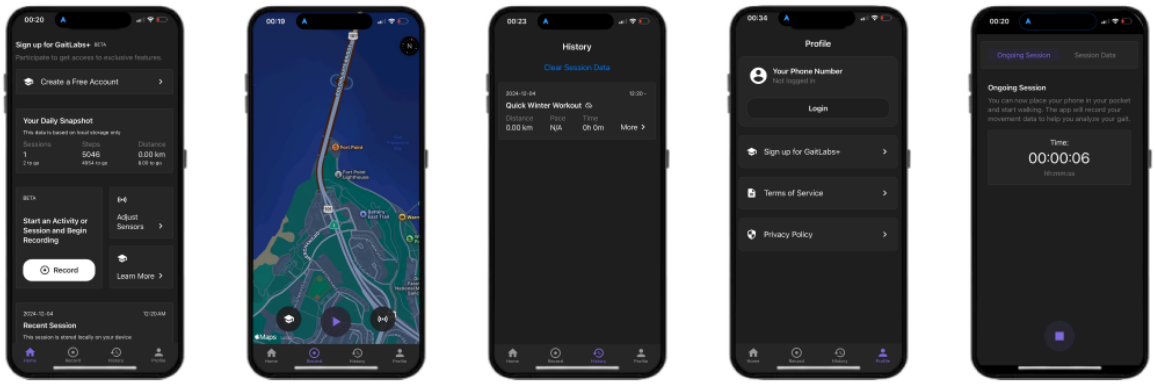


Note: Raw accelerometer, gyroscope, and magnetometer readings during walking trials on the Samsung Galaxy S10e (2019), iPhone 14 Pro (2022), and MetaMotionS IMU showing x, y, z components and the resultant magnitude. Pauses at the start and end of the activity were used to isolate walking motion from noise artifacts.

Figure 5.6 Sensor Data During 10-Meter Walking Trials with Smartphone in Pocket

The app was designed with features that encourage user engagement without feeling manipulative. For instance, the application is fully functional both offline and online, but cloud storage is only available to users who create an account. This approach balances financial considerations with user incentives. To further encourage account creation, a banner is displayed on the homepage, linking to a call-to-action (CTA) for signing up (Figure 5.7; Home Tab). The banner disappears once an account is created, providing a subtle but effective "negative reinforcement" by removing a persistent element in

exchange for user engagement. These design choices aim to strike a balance between functionality, user experience, and research goals.



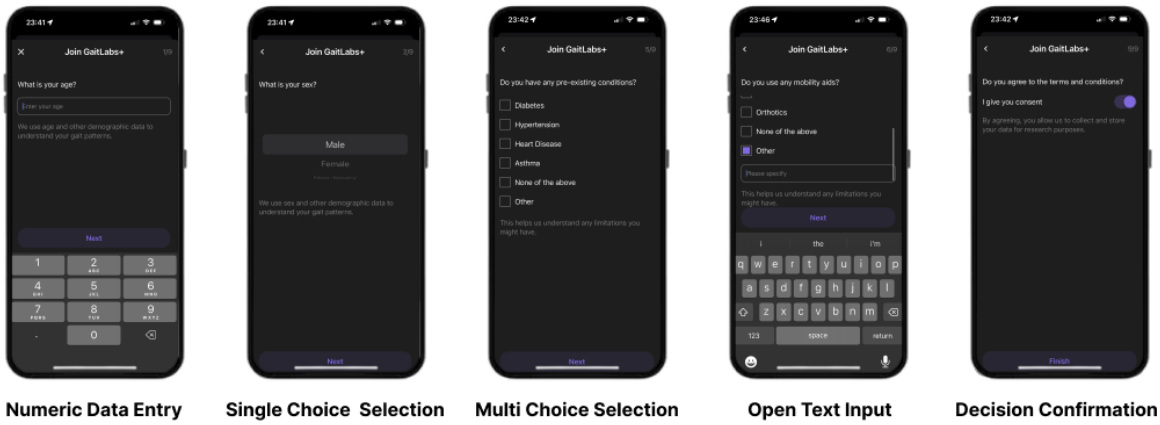
Note: Displays key application screens, including the Home Tab, Record Tab, History Tab, Profile Tab, and Recording Screen

Figure 5.7 Application Screens

5.4.4 Interface for Heterogeneous Health Data: To support the collection of health data for future studies, the app included input fields designed to record information about comorbidities, other health factors, and participant consent (Figure 5.8). While open-ended questions could provide flexibility, they risked introducing significant data heterogeneity, as varied responses would be difficult to standardize for modeling or reporting.

To address this, the app was designed to offer predefined response options for questions where standardization was critical. For instance, in the case of walking aids, participants could select from common options, with additional choices for "None of the above" and "Other" to maintain flexibility while promoting consistency. Other fields, such as age, weight, or height, were designed to use a numeric input interface with a numpad and appropriate range limitations to ensure accuracy and prevent erroneous entries.

5.4.5 Strengths and Limitations: Our system demonstrated several strengths, including cross-platform compatibility through React Native, scalable architecture for managing high-frequency sensor data, and a user-friendly interface designed for accessibility and usability. Offline functionality ensures reliability of data collection in real-world environments, and the integration of IMU and GPS data offers novel analyses, such as gait assessment and location-based insights.



Note: Illustrates intuitive user experience for input and consent flow, featuring Numeric Data Entry, Choice Selections, Open Text Input, and Decision Confirmation.

Figure 5.8 User Info and Consent Flow

However, there were notable limitations. Hardware and software variability presented challenges, particularly with inconsistent sampling rates for accelerometer and gyroscope sensors. Additionally, relying on a single device sensor, while practical, may lack the resolution required for certain outcomes, such as step width, which typically requires multiple sensor placements in biomechanical studies.

Finally, although our testing was limited to a small number of development devices and personnel, the wide variety of smartphones on the market and the diversity in their usage and placement (e.g., pockets, bags) introduce substantial heterogeneity. While modern smartphones are computationally powerful, fully harnessing their potential will require advanced models capable of accommodating these variations.

5.5 Conclusion

This chapter outlined the design, implementation, and validation of a smartphone-based data collection system for mobility tracking. By integrating IMU and GPS sensors with scalable data storage and an accessible user interface, the system was designed to facilitate health data collection across diverse contexts. However, the inherent heterogeneity in smartphone hardware, user behavior, and environmental conditions presents significant challenges, particularly for standardization and modeling. Addressing this complexity requires advanced analytical approaches, as it is not feasible to rely solely on algorithmic programming to account for such variability. This chapter sets the stage for the next, where human activity recognition (HAR) techniques are introduced. These methods leverage deep learning to manage the heterogeneity observed in sensor data, enabling robust and scalable solutions for personalized health monitoring.

6. Deep Learning Models with Smartphone Sensor Data for Human Activity Recognition: Literature Review and Case Study

In Chapter 5, we demonstrated how smartphone devices serve as practical and efficient tools for collecting movement and activity data. Their real-time capabilities, scalability, and ease of use make them valuable for various applications. However, challenges remain, particularly those stemming from sensor heterogeneity. Variations in sensor placement, smartphone models, and operating systems (e.g., iOS vs. Android) can significantly affect data quality and consistency.

Chapter 6 builds on this foundation by addressing these challenges and expanding the scope to tackle the issue of manual activity labeling. First, we conduct a literature review of automated activity labeling technologies designed for wearable devices. Then, we pretrain and compare three deep learning models—densely connected feed-forward networks, Long Short-Term Memory (LSTM) networks, and Gated Recurrent Units (GRU)—to evaluate their effectiveness in low-cost activity classification. This chapter serves as a bridge to Chapter 7, where we will integrate insights from various domains, including gait analytics, systematic reviews, and Human Activity Recognition (HAR) technologies,

6.1 Introduction

6.1.1 The Growing Potential of Smartphone Sensors: The increasing proliferation of smartphones—approximately 5 billion devices globally in 2024 alone—has ushered in unprecedented opportunities for human activity recognition (HAR) through embedded sensor technology [227]. These devices, equipped with accelerometers, gyroscopes, magnetometers, GPS, and other advanced sensors, enable the continuous collection of detailed data on human movement and behavior. Smartphones' ubiquity, computational power, and accessibility have transformed them from mere communication tools into robust platforms for health monitoring and activity tracking.

Their accessibility also makes them ideal for large-scale health initiatives. Unlike niche wearable devices, smartphones are globally pervasive and widely accepted across age groups. This ubiquity minimizes barriers to adoption and allows researchers to leverage a familiar platform for collecting and analyzing sensor data. Their potential for HAR is further amplified by advancements in computational techniques and AI algorithms, which allow smartphones to process complex sensor data efficiently in real time.

6.1.2 Challenges in Human Activity Recognition: Despite their promise, smartphones face several challenges in HAR. Sensor heterogeneity—variations in smartphone models, operating systems, and sensor placements—complicates data consistency and quality. These differences can result in noise, missing data, and biased results, particularly in diverse real-world environments. Moreover, manual activity labeling for training HAR models remains a significant bottleneck due to its labor-intensive nature and susceptibility to subjectivity [228], [229].

To overcome these challenges, researchers have explored automated annotation techniques and robust preprocessing methods. Techniques like sensor fusion and orientation-invariant preprocessing have shown promise in mitigating inconsistencies caused by heterogeneous devices. Additionally, leveraging semi-supervised and transfer learning approaches has the potential to scale HAR applications, addressing data scarcity while improving model generalizability across diverse datasets.

6.1.3 Innovations in Data Annotation and Classification: To address the bottleneck of manual data annotation, we conducted a literature review of automated annotation techniques, exploring methods such as active learning, transfer learning, and sensor fusion. These approaches reduce the reliance on human annotators, enabling scalable and efficient data labeling processes. Automated annotation systems are particularly valuable for large-scale studies in free-living conditions, where traditional manual labeling is impractical [229].

In addition to literature review, we evaluated three deep learning architectures—densely connected feed-forward networks (FFN), Long Short-Term Memory (LSTM) networks [230], and Gated Recurrent Units (GRU) [231]—for activity classification using publicly available HAR datasets. Pre-trained models may offer a cost-effective solution for activity classification and serve as a strong baseline for integrating more contextual features, such as GPS data, to enhance resolution and accuracy [232].

These innovations align with broader efforts to advance AI-driven healthcare solutions. By addressing critical gaps in annotation and classification, this work contributes to the development of scalable, real-time HAR systems capable of supporting preventative health monitoring and disease management.

6.2 Methodology

6.2.1 Database Search of Automated HAR Labelling: To supplement our results, we performed a literature review of automated labelling methods for the purpose of HAR, following PRISMA guidelines and registered with PROSPERO (registration number CRD42024538078). Database search was conducted in April 2024, covering studies published within the last ten years. We searched the following

databases: Embase, CENTRAL, PubMed, Web of Science, and CINAHL and utilized a comprehensive set of keywords and MeSH terms as outlined in Table 6.1.

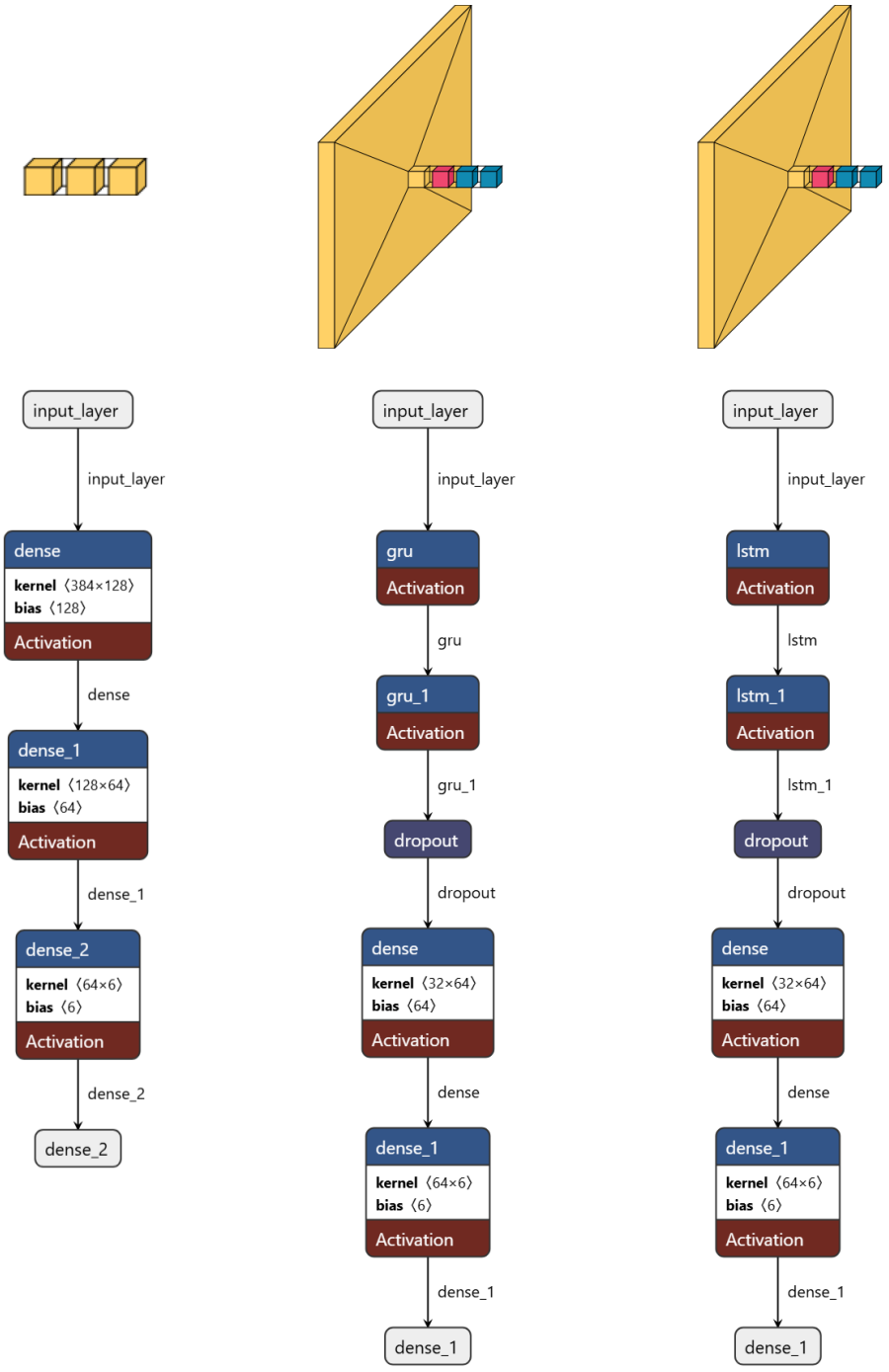
Table 6.1 Keywords Used in Database Search

Category	Keywords
Wearable Sensor Technology	Wearable*, Sensor*, Accelerometer, Gyroscope, Magnetometer, Barometer, Smartphone, Smartwatch, “Inertial Sensor”, "Inertial Measurement Unit (IMU)", "Global Positioning System (GPS)", “Wearable Sensor”, "Wearable Computer", “Activity Tracker”, "Fitness Tracker", "Body Area Network"
Human Activity Recognition (HAR)	Activity*, Task*, Recognition*, Classification*, “Human Activities”, "Human Activity Recognition (HAR)", “Daily Life Activity”, "Activities of Daily Living (ADL)", "Physical Activity", "Motor Activity", "Behavioral Monitoring"
Data Preprocessing and Feature Extraction	Denoising, Normalization, Segmentation, Pre-processing, "Sensor fusion", "Time-Frequency Analysis", "Signal Processing", "Fourier Transform", "Feature Extraction"
Labelling	Automatic*, Label*, Labelling*, Annotate*, Annotation*, "Automatic Annotation", "Data Annotation", "Crowdsourcing"
Machine Learning Techniques and Algorithms	Supervised*, Self-Supervised*, Unsupervised*, Reinforcement*,"Machine Learning", “Artificial Intelligence” “Artificial Neural Network”, “Supervised Machine Learning”, “Unsupervised Machine Learning”, "Deep Learning", "Decision tree", "K-nearest neighbors", "Support Vector Machine", "Convolutional Neural Network", “Recurrent Neural Network”, "Semi-Supervised Machine Learning", “Automated Pattern Recognition”, "Self-training"

Note: In the database search, categories were combined using the AND operator, and keywords within each category were combined using the OR operator.

While the methodology was informed by systematic review principles, certain elements, such as the PRISMA flow diagram, risk-of-bias assessment, and quality scoring, were omitted to better align with the objectives and scope of the project.

6.2.2 Screening and Data Extraction: Title and abstract screening was automated through our custom web application using the OpenAI API with the GPT-3.5 Turbo model. This application used a tailored prompt template to classify studies as 'include' or 'exclude'. Subsequently, full texts of included studies were segmented into 1000-token chunks with 100-token overlaps and indexed using Hierarchical Navigable Small World (HNSW) vector stores for efficient querying. Data was exported in .csv format.



Note: The figure illustrates three neural network architectures: a densely connected network (FFN), a gated recurrent unit (GRU) network, and a long short-term memory (LSTM) network. The FFN model includes multiple fully connected layers with specified kernel and bias sizes. The GRU and LSTM models incorporate recurrent layers to handle sequential data, with dropout layers for regularization

Figure 6.1 Deep Learning Model Architectures

To simulate the systematic review guidelines of having two independent reviewers, we conducted two separate runs of the classification process using the same model. The first run used the standard

prompt, while the second run incorporated a modified prompt that asked the model to provide an explanation for its classification. Additionally, we validated the screening process using an ordinal prompt, instructing the model to rate each study on a 1–10 scale to provide more nuanced insights. Conflicts and final inclusion decisions were made through human judgment.

6.2.3 Description of Preliminary Models and Model Architecture: We compared three deep learning models: a densely connected feed-forward network (FFN), gated recurrent unit network (GRU) [231], and long short-term memory network (LSTM) [230]. FFNs were included as a control; they process data in a single direction and are effective for pattern recognition but lack mechanisms to model temporal dependencies. GRUs and LSTMs were chosen for their feedback loops, which allow them to capture temporal relationships, making them ideal for activity recognition tasks with sensor data. The architectures of the chosen models are shown in Figure 6.1.

6.2.4 Training Dataset Description and Preprocessing: We utilized the Heterogeneity Activity Recognition dataset from the University of California, Irvine Machine Learning Repository [233]. The dataset comprised six labeled activities: Biking, Sitting, Standing, Walking, Stair Up, and Stair Down. It contained maximum sampling frequency readings from the tri-axial accelerometer and gyroscope of eight different smartphone devices. The inclusion of data from multiple devices was particularly relevant in our circumstance, where variations in hardware and sensor placement can significantly impact model performance.

Summary statistics (mean, standard deviation, quartiles) were computed for each sensor and visualized as frequency histograms. Data segmentation was then performed using a sliding window approach with a window size of 128 data points and a step size of 64. The segmented data was randomly shuffled and split into training (70%), validation (15%), and test (15%) sets. To standardize the data and account for variations between devices, Z-score normalization was applied using the formula:

$$Z_i = \frac{x_i - \underline{x}}{\sigma}$$

where Z_i represents the normalized value, x_i represents the sensor reading at each index, $\underline{x}$ represents the mean sample sensor value, and σ represents the sample standard deviation. Activities were one-hot encoded for categorical representation.

6.2.5 Model Training: Models were trained separately on accelerometer and gyroscope data to evaluate their individual contributions to activity recognition. We employed categorical cross-entropy as the loss function and ReLU activation functions for intermediate layers. Training was conducted over 20 epochs

with a batch size of 64, using the Adam optimizer with the default learning rate of 0.001 [234]. Early stopping and dropout regularization [235] was assessed in both recurrent models with a dropout rate of 0.5 to mitigate overfitting and prevent the network from relying too heavily on any one feature.

To further optimize performance, we experimented with various architectures, including CNN-LSTM and CNN-GRU hybrids, batch normalization [236], and adjustments to the number of units and layers in the recurrent models. However, these configurations did not yield significant improvements over the simpler architectures and were thus excluded from the final results for brevity.

6.2.6 Supplementary Dataset Preparation, Evaluation Metrics and Analysis: New supplementary activity data was collected from two researchers performing the same activities as those in the training dataset. Each activity was performed for 15-30 seconds and subsequently combined, with the first and last 5 seconds of data removed to ensure consistency for automated labeling. Data collection utilized a Samsung Galaxy S10e (2019), iPhone 14 Pro (2022), and MetaMotion IMU [226] at a sampling rate of 50 Hz. To account for variations in sensor output, all readings were converted to common units prior to analysis.

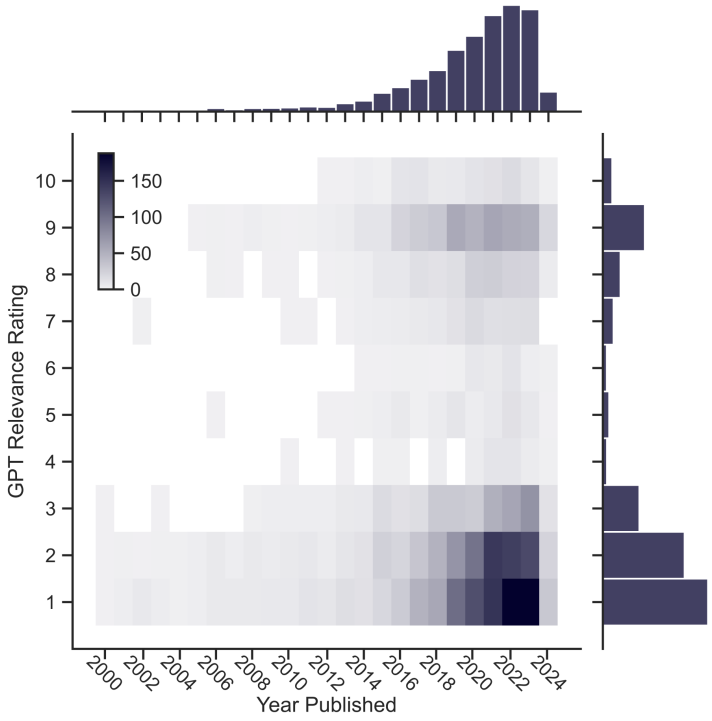
To address differences in sampling rates, models were retrained on a subsampled version of the original dataset. Pre-trained model predictions were applied to this supplementary dataset for labeling, with ground truth determined through manual annotation. Confusion matrices were analyzed to identify common misclassifications, including overlaps between similar activities. Additionally, the performance of the pre-trained model was evaluated on its own test set as well as the original dataset, with a focus on F1-score as the primary metric, while accuracy, precision, and recall were also assessed.

6.3 Results

6.3.1 Literature Review Results: The results of the initial database search are summarized in Figure 6.2. The analysis highlights a growing interest in HAR and wearable technologies, particularly within the last five years. This trend is visualized in the accompanying heatmap, which illustrates a significant rise in relevant publications and research activity over this period.

Particularly over the last five years, studies have expanded beyond simple daily activity monitoring to include specialized medical and occupational tasks, drawing on works that date back more than a decade. In particular, there is growing emphasis on continuously monitoring free-living conditions for individuals with Parkinson's disease [237], [238], [239], [240], epilepsy [241], frailty [242], autism

spectrum disorder stereotypies [243], as well as on designing semi- and fully-automated annotations for large-scale datasets [244], [245], [246], [247].



Note: Depicts a heatmap showing the distribution of GPT relevance ratings against the year of publication. Darker regions indicate a higher density of articles with greater relevance scores, suggesting trends over time. Marginal histograms on the top and right display the frequency of publications per year and the distribution of relevance ratings, respectively.

Figure 6.2 Year of Publication Plotted Against Article Topic Relevance

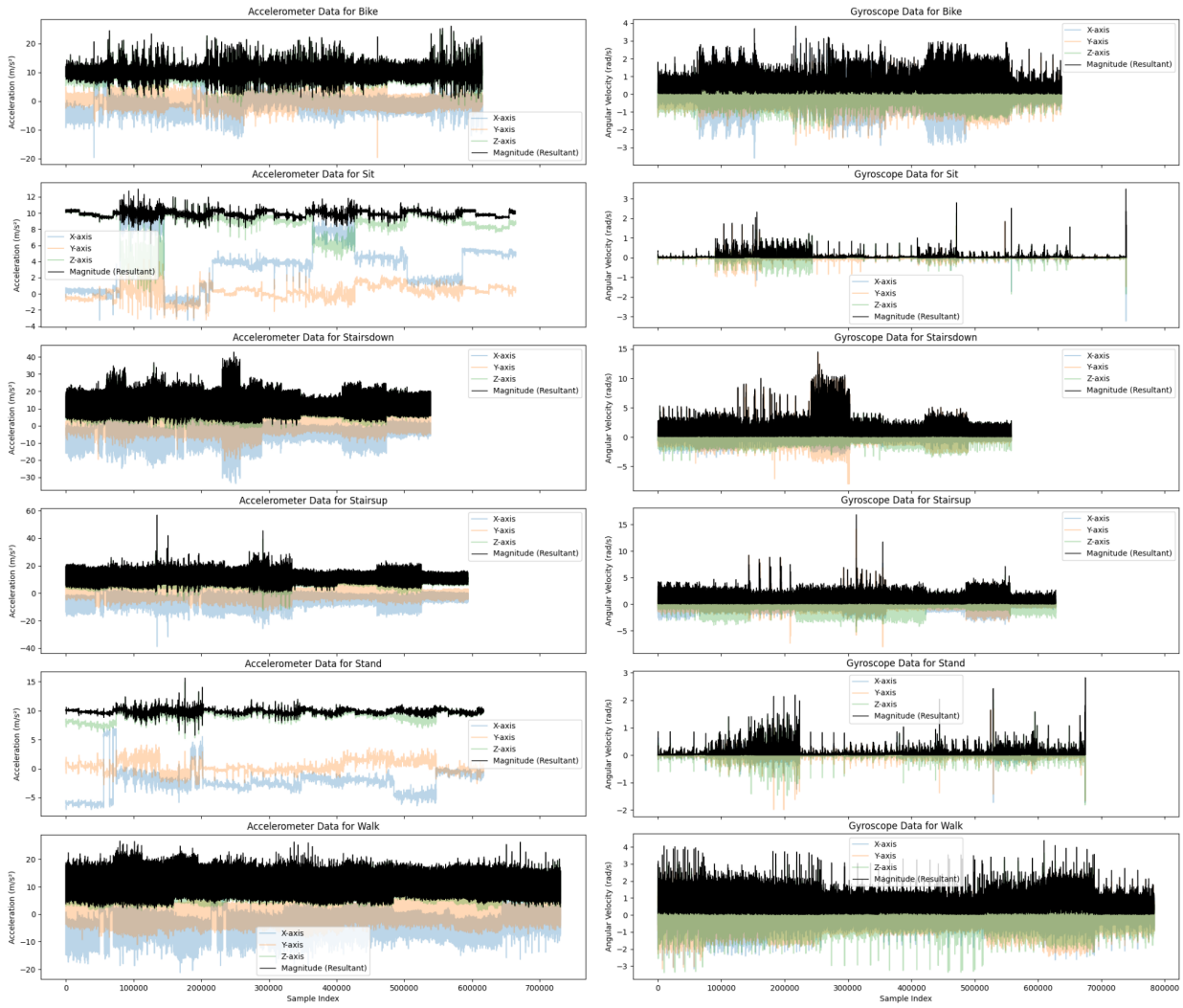
Recent work reveals a diverse array of devices and sensor types employed to capture human motion and physiological states, ranging from simple accelerometers to multimodal suites that include gyroscopes, magnetometers, and even electrocardiogram (ECG) and electromyography (EMG) sensors [241], [248], [249], [250], [251], [252], [253]. Some rely on smartphones with embedded IMUs [247], [254], [255], [256] while others leverage wrist-worn devices or smartwatches (e.g., Apple Watch or custom wristbands) [238], [239], [242], [257]. Meanwhile, other specialized setups attach IMUs at the waist, ankles, or shoes for gait analysis [237], [258], [259], and construction or rehabilitative scenarios may involve thigh- and calf-mounted sensors [249], [260].

In terms of activities studied, most investigations included a baseline set of daily living tasks (e.g., walking, running, sitting, and standing) as fundamental benchmarks for HAR [244], [246], [247], [248], [250], [255], [261], [262], [263]. Many have also focused on stair climbing to evaluate more challenging transitions [247], [248], [250], [254], [255], [256], while others included specialized tasks

such as race walking [256], industrial tasks like kneeling or picking up loads [249], [260], or compensatory balance responses for fall-risk assessment [264]. In each scenario, attention is given to collecting realistic data—often indoors for controlled assessments [257], or in free-living and outdoor contexts to improve ecological validity [237], [239], [265].

From a modeling and automatic labeling perspective, recent literature demonstrates a significant push toward deep learning, including CNN+LSTM hybrids [248], [249], [266], transformers [267], and self-supervised frameworks such as contrastive learning or masked autoencoders [246], [254]. Classical ML algorithms remain prevalent, particularly SVMs, random forests, and k-nearest neighbors, either for baseline comparisons or in hybrid pipelines [240], [247], [256], [268], [269]. Moreover, semi-supervised approaches (e.g., adversarial autoencoders) [245] and unsupervised methods (e.g., clustering or hidden Markov model regression) [255], [270] address the costly challenge of manual annotations, sometimes leveraging video-based synthetic data [244]. Additional feature extraction methods ranged from wavelet transforms [248], and Fourier approximations [240] to more direct end-to-end learned features [271]. Novel feature-extraction techniques have also come to light, such as symbolic approximations [240], contrastive learning [254], and codebook-based approaches [272].

6.3.2 Training Dataset Description: The training dataset consisted of raw accelerometer and gyroscope sensor readings collected from smartphones, categorized by labeled activities such as walking, sitting, and biking (Figure 6.3). The data consisted of tri-axial accelerometer and gyroscope readings collected at ~159 Hz (accelerometer) and ~170 Hz (gyroscope) from eight different smartphone devices.



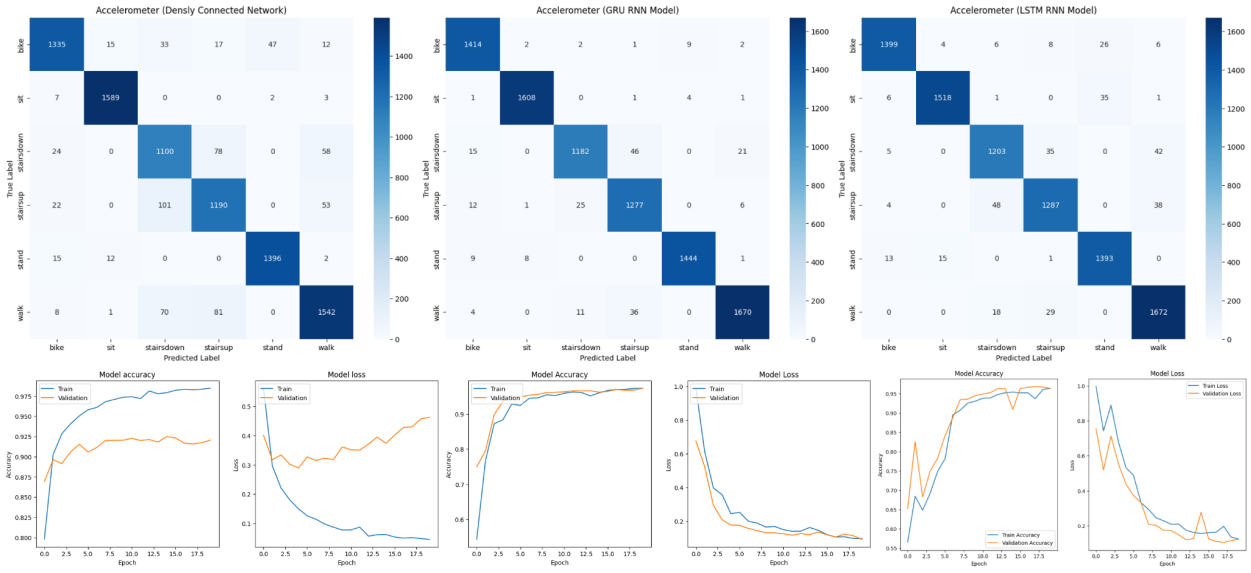
Note: Resultant and raw accelerometer and gyroscope sensor data collected from smartphone devices, organized by labeled activities. The data, sourced from the UCI Machine Learning Repository, provides a basis for activity recognition.

Figure 6.3 Resultant Smartphone Sensor Data Sorted by Activity

6.3.3 Model Performance Across Accelerometer and Gyroscope Data: Figure 6.4 illustrates the performance of three models trained on accelerometer data to classify activities. The performance was generally good, as indicated by the strong diagonal patterns in the confusion matrix heatmaps. However, overfitting was observed in the DNN, as evident from the divergence between training and validation accuracy curves in the lower panel. The LSTM and GRU models demonstrated superior performance, effectively capturing temporal dependencies with higher accuracy and better generalization compared to the other model.

Across all models, classification performance was particularly strong for static activities, such as "Sit" (F1-score: 0.98-0.99) and "Stand" (F1-score: 0.97-0.99). Temporal activities, such as "Stairs Up"

(F1-score; DNN: 0.87, GRU: 0.95, LSTM: 0.94) and "Stairs Down" (F1-score; DNN: 0.86, GRU: 0.95, LSTM: 0.94), posed greater challenges, with the DNN exhibiting noticeable limitations due to its inability to model sequential patterns effectively. GRU achieved the highest overall accuracy at 0.98, followed by LSTM at 0.96 and DNN at 0.92, highlighting the advantages of recurrent architectures in leveraging temporal dependencies for accelerometer data.

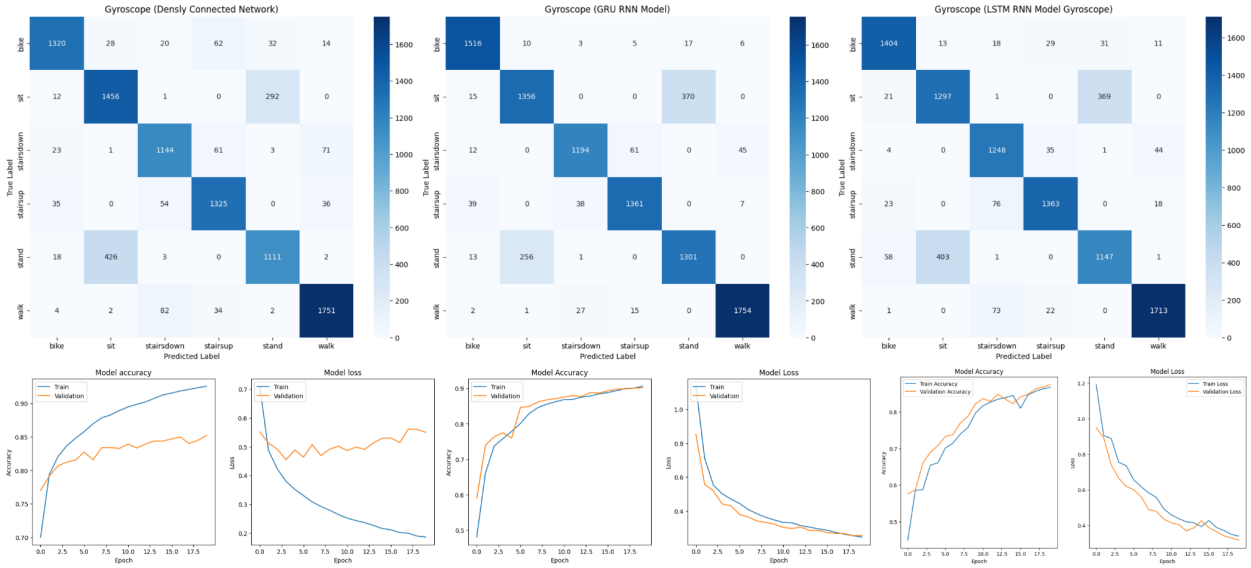


Note: Performance of three models (Dense Neural Network, GRU, and LSTM) for classifying accelerometer data, shown through confusion matrices and performance metrics. The top half shows confusion matrices, where darker diagonal cells indicate higher classification accuracy across activities. The bottom half illustrates training and validation loss, as well as accuracy trends, across 20 epochs for each model.

Figure 6.4 Comparison of Accelerometer Based Performance Using Confusion Matrices and Model Metrics

Similarly, Figure 6.5 presents the performance of the same models trained on gyroscope data. While the confusion matrices show generally strong performance, the gyroscope-trained models exhibited more confusion between sitting and standing activities compared to the accelerometer models. As with the accelerometer data, overfitting was observed in the DNN, while the LSTM and GRU models again demonstrated superior generalization and more stable learning patterns across 20 epochs.

Across all models, gyroscope data classification showed strong performance overall but slightly lower metrics compared to accelerometer data, particularly for static activities like "Sit" (F1-score: 0.76-0.81) and "Stand" (F1-score: 0.73-0.80). Temporal activities like "Stairs Up" (F1-score; DNN: 0.90, GRU: 0.94, LSTM: 0.93) and "Stairs Down" (F1-score; DNN: 0.88, GRU: 0.93, LSTM: 0.91) demonstrated better classification compared to static activities. Overall, the GRU achieved the highest accuracy at 0.90, followed by the LSTM at 0.87, and the DNN at 0.86.

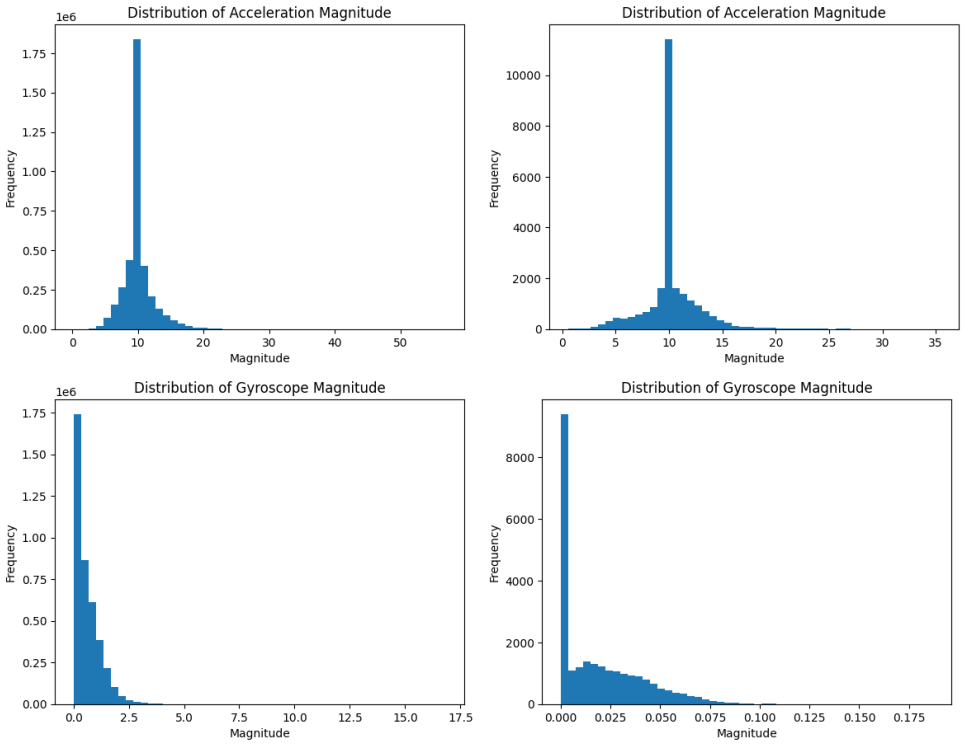


Note: Performance of three models (Dense Neural Network, GRU, and LSTM) for classifying gyroscope data, shown through confusion matrices and performance metrics. The top half shows confusion matrices, where darker diagonal cells indicate higher classification accuracy across activities. The bottom half illustrates training and validation loss, as well as accuracy trends, across 20 epochs for each model.

Figure 6.5 Comparison of Gyroscope Based Performance Using Confusion Matrices and Model Metrics

6.3.4 Supplementary Dataset Description: Figure 6.6 compares the distributions of accelerometer and gyroscope magnitude readings between the training dataset and the Supplementary Dataset. The accelerometer data exhibited a normal distribution in both datasets, while the gyroscope data showed a distinct right skew. A notable discrepancy was observed in the gyroscope magnitude: the training dataset reached up to 2.5, whereas the Supplementary Dataset only extended to 0.075. This difference likely reflects variations in data collection or sensor behavior between the two datasets, potentially influencing model performance when applied to the Supplementary Dataset.

6.3.5 Supplementary Dataset Results: Figure 6.7 presents the performance of trained models when applied to the Supplementary Dataset, using resultant accelerometer, gyroscope, and magnetometer signals segmented by activities. The confusion matrices in the bottom rows reveal poor performance across all models for both accelerometer and gyroscope data. The highest F1-score achieved was 0.66 for the accelerometer LSTM model on the "sit" activity, while overall accuracy ranged from 0.23 to 0.33. These results indicate significant challenges in generalizing the trained models to the new dataset, likely due to discrepancies in data distributions or sensor characteristics.



Note: Shows the distributions of accelerometer and gyroscope magnitude readings for both training (left) and supplementary (right) datasets. The top row illustrates the distribution of acceleration magnitude, while the bottom row shows gyroscope magnitude distributions.

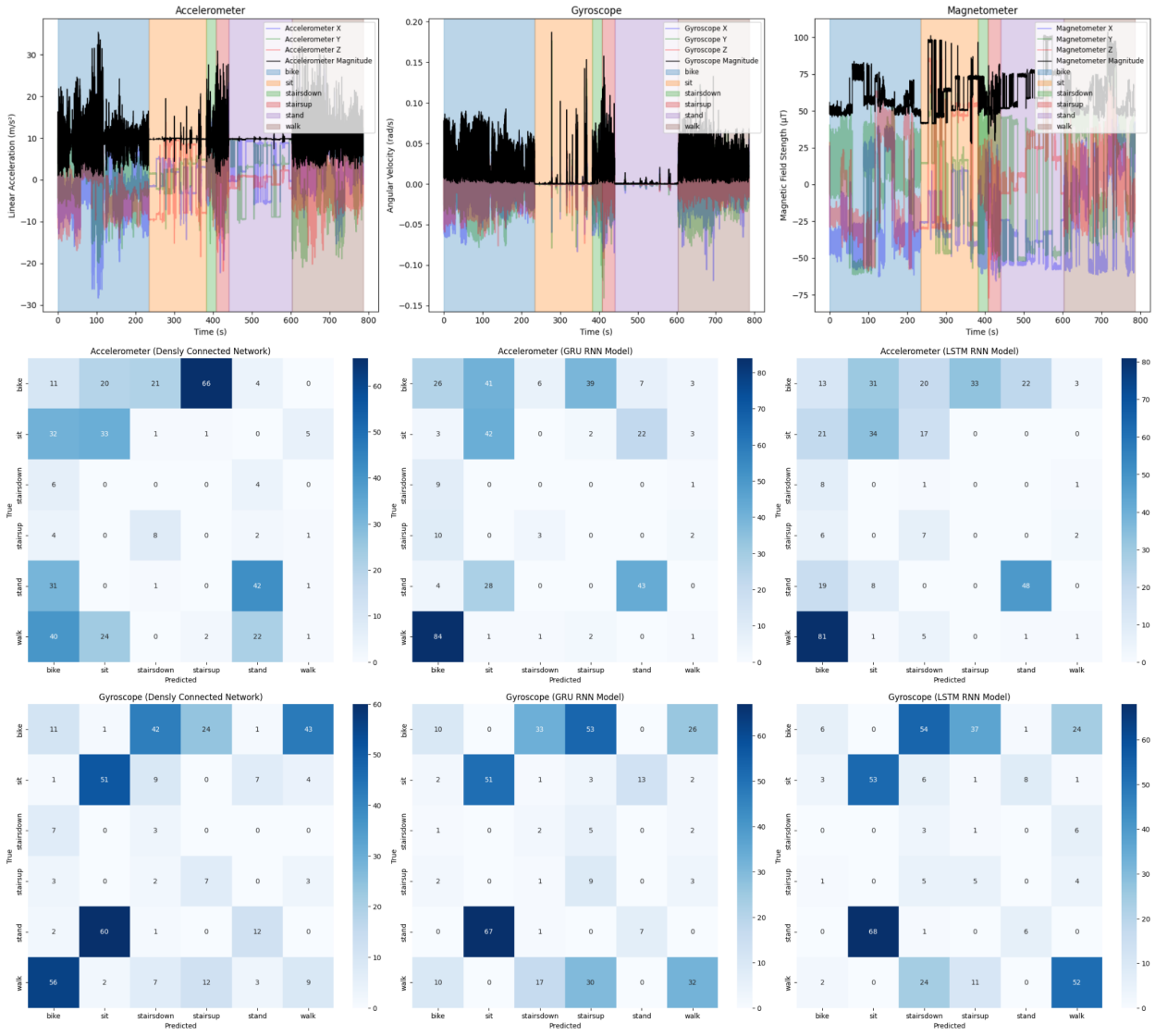
Figure 6.6 Training and Supplementary Distributions of Accelerometer and Gyroscope Sensor Readings

6.4 Discussion

6.4.1 Generalization Challenges in HAR Models: Despite recent innovations, challenges remain in achieving robust HAR across diverse users, contexts, and sensor noise levels [230,231,234,253]. Variations in device placement, inter-person motion differences, and limited labeled datasets continue to hamper model generalization [239,256,258,266]. Clinical populations, such as individuals with Parkinson's disease, COPD, or autism, introduce further variability in motor profiles that can confound algorithm performance [233,236,258]. Additionally, the move toward real-world "free-living" monitoring imposes additional constraints on battery life, comfort, and compliance [232,267]. Nevertheless, ongoing efforts that emphasize multi-sensor fusion, self-supervised or semi-supervised learning, and carefully curated benchmark datasets for open comparison are expected to drive more reliable and scalable HAR solutions for both healthy and clinical populations [233,235,241,246].

The results from this work highlight the significant challenges of generalizing deep learning models for HAR across diverse datasets. While the trained models demonstrated strong performance on the original training and test datasets, with high classification accuracy and stable training curves, their

performance on the unseen Supplementary Dataset was notably poor. The highest F1-score achieved on the Supplementary Dataset was only 0.66 (accelerometer LSTM for "sit"), and total accuracy ranged from 0.23 to 0.33 across models. These findings suggest a clear domain shift between the original training data and the Supplementary Dataset, which is a well-documented problem in HAR research.



Note: The top row shows resultant accelerometer, gyroscope, and magnetometer signals segmented by activities. The bottom rows include confusion matrices for Dense Neural Network, GRU, and LSTM models applied to accelerometer and gyroscope data.

Figure 6.7 Test Set Resultant Sensor Readings and Confusion Matrices

6.4.2 Implications for HAR Research and Applications: The poor performance of trained models on the Supplementary Dataset emphasizes the critical need for addressing domain shift in HAR research. These findings align with broader challenges in the field, where variations in sensor hardware, data collection

protocols, and participant behaviors often hinder generalization to unseen environments. For real-world applications such as health monitoring or activity tracking, deploying HAR systems trained on limited datasets without thorough validation poses significant risks. Misclassifications in critical activities, such as "stairsdown," could have safety implications in healthcare or assistive technologies.

At the same time, this study highlights the strong performance of models trained on raw sensor signals, even without explicit feature engineering. This demonstrates the potential of deep learning models to learn meaningful patterns directly from raw data, reducing reliance on manual preprocessing steps. However, while the models performed well on their original test sets, their inability to generalize effectively to the Supplementary Dataset raises important questions about the role of domain shift and data representation in HAR research.

Our additional experimentation, including hybrid models (e.g., CNN-LSTM, CNN-GRU), batch normalization, layer normalization, hyperparameter tuning, and regularization techniques, further reinforces these findings. Despite these efforts, no significant improvements in performance were observed. This suggests that the primary limitation lies not in model architecture but in the variability and inconsistency of the data itself. Addressing these domain shifts, rather than solely refining model architectures, may hold the key to improving generalization in HAR systems.

6.4.3 Dataset Discrepancies and Study Limitations: A closer examination of the dataset distributions reveals critical differences between the training and Supplementary Datasets that likely contributed to poor generalization. Specifically, while the accelerometer data showed consistent normal distributions in both datasets, the gyroscope data displayed a pronounced right skew in the Supplementary Dataset, with magnitudes only reaching 0.075 compared to 2.5 in the training data.

However, the poor performance of accelerometer-based predictions suggests that these discrepancies alone do not fully account for the observed generalization issues. If the mismatch in gyroscope data were the sole factor, we would expect the accelerometer models to perform significantly better on the Supplementary Dataset. Instead, both sensor types demonstrated poor generalization, indicating that domain shift extends beyond individual sensor discrepancies and encompasses broader issues in data representation and collection protocols.

Finally, the lack of specific details about the activities in the training dataset—such as differences in bicycle type (e.g., stationary bike), staircase design, and phone placement—combined with the need to subsample sensor readings for the Supplementary Dataset (50 Hz compared to ~159–170 Hz in the training data), may have further impacted the models' ability to extract meaningful temporal patterns.

6.4.4 Future Directions: Improving the generalizability of HAR models requires addressing the challenges posed by domain shift and variability in sensor data. One promising avenue is the use of domain adaptation techniques, which allow models to adjust to the specific characteristics of new datasets. For example, fine-tuning pre-trained models on small subsets of new data could help mitigate domain shift by aligning the model with the features and distributions of the Supplementary Dataset. Complementary to this, data augmentation techniques, such as adding noise, simulating variations in device orientation, or scaling sensor readings, could improve model robustness to sensor variability and better prepare models for real-world conditions.

In addition to these model-level strategies, future efforts should focus on the collection of larger and more diverse datasets. Datasets that incorporate a variety of devices, environments, and activity scenarios will enable the development of models capable of handling the complex variability encountered in real-world applications. Publicly available datasets could serve as valuable resources for transfer learning, where pre-trained models are fine-tuned on domain-specific datasets to improve performance under unseen conditions.

Finally, integrating additional contextual features, such as GPS, environmental data, or user metadata, may enhance activity recognition by providing complementary information to resolve ambiguities in activity transitions. These approaches, combined with robust validation across diverse settings, will be crucial for advancing HAR systems toward reliable deployment in real-world applications.

6.5 Conclusion

In this chapter, we explored the potential of deep learning models for Human Activity Recognition (HAR) using smartphone sensor data, addressing challenges such as manual labeling and domain shift. While recurrent models like GRU and LSTM demonstrated strong performance on the training dataset, achieving high F1-scores without explicit feature engineering, their poor generalization to the Supplementary Dataset highlighted the critical impact of domain variability. Key issues included discrepancies in sensor distributions, sampling rates, and activity protocols, which architectural adjustments alone could not resolve. These findings emphasize the need for domain adaptation, data augmentation, and diverse dataset collection to improve robustness. By addressing these challenges, future HAR systems can achieve greater reliability and scalability for real-world applications.

7. Conclusion and Future Directions

This thesis has examined the integration of advanced computational tools and artificial intelligence (AI) across multiple healthcare domains. The overarching objective was to address fundamental challenges in (1) identifying distinct gait signatures associated with dementia subtypes, (2) accelerating and improving the systematic review process through large language models (LLMs), and (3) leveraging smartphone-based sensors to facilitate human activity recognition (HAR) with minimal manual annotation and reduced sensor heterogeneity. Taken together, these contributions highlight the potential of AI and digital health technologies to enhance diagnosis, research translation, and long-term monitoring within healthcare.

Although gait analysis, systematic reviews, and HAR might initially appear as distinct areas, this thesis has illustrated the synergistic potential of integrating AI-driven methods across these domains. The improved diagnostic precision in dementia research could inform targeted interventions, and tools for automated screening of research literature can expedite the adoption of new scientific evidence into clinical practice. Meanwhile, smartphone-based HAR solutions can serve as frontline data-collection platforms, providing continuous, real-world measurements that complement clinical insights. Taken together, these contributions form a feedback loop: as evidence accumulates and clinical insights evolve, AI-driven systematic reviews can incorporate fresh data from new sensor-based studies, fostering a continuously improving ecosystem of research, practice, and patient-centered care.

7.1 Summary of Key Findings

7.1.1 Gait Signatures in Dementia: Chapter 2 examined gait parameters, such as stride length, walking speed, and stance time, across Alzheimer's disease (AD), Lewy body dementia (LBD), frontotemporal dementia (FTD), and vascular dementia (VaD). LBD often showed the most pronounced disruptions, while FTD presented milder variations. These findings suggest that gait evaluation might serve as a non-invasive marker for early and subtype-focused interventions, although inconsistencies in measurement protocols can complicate interpretation.

Efforts to standardize data collection and apply strategies like individual participant data analyses may address this variability [273]. More unified approaches to instrument calibration and study design could clarify which gait metrics reliably track disease onset or progression. By refining these protocols, future research may yield more robust conclusions about the clinical role of gait-based indicators in dementia care.

7.1.2 Systematic Review Automation with LLMs: Chapters 3 and 4 explored large language models as a way to streamline screening in systematic reviews, revealing a notable reduction in manual workload when ordinal prompting was used. Although these automated methods helped balance sensitivity and specificity, human oversight remained essential for identifying critical studies that models might miss [274]. This highlights the importance of rigorous validation metrics and transparent reporting of model details. Recent work has also shown that machine learning contributes to diverse clinical tasks, from diagnostic image analysis to broader data mining [275], [276]. Such a setup could further expand the scope and reliability of AI-driven systematic reviews. Moving toward semi- or fully automated workflows will depend on open collaboration, reproducible results, and continued refinement of both model interpretability and user acceptance.

7.1.3 Smartphone-Based Human Activity Recognition: Chapters 5 and 6 introduced a smartphone-based system that classifies daily activities, such as walking, sitting, and standing, using built-in inertial sensors. Controlled tests showed promising accuracy and a practical route toward continuous monitoring of mobility and possible early signs of functional decline. However, sensor heterogeneity and domain shift can reduce performance when transitioning from a laboratory to real-world scenarios.

Developers can mitigate these challenges through standardized calibration, data augmentation, and on-device processing that preserves user privacy. These adaptations could help align the technology with clinical demands, particularly for older adults or those with chronic conditions. By refining smartphone-based monitoring and validating it under diverse field conditions, the approach may become a scalable option for real-time health tracking.

7.2 Limitations and Challenges

7.2.1 Gait and Meta-Analysis Challenges: Analyses in Chapter 2 showed that diverse instruments (inertial measurement units, camera-based systems, force plates) and uneven protocols often yield high I^2 values, complicating the interpretation of pooled effect sizes. Although I^2 is widely used, it does not always reflect the actual range of effects and must be approached with caution [277]. Reliance on techniques like median-imputation can further skew comparisons between dementia subtypes if data are missing or inconsistently reported. More uniform procedures for gait measurement, including standardized walking distances and environmental conditions, could reduce this variability. Meta-regression and individual participant data methods [273] may also help clarify whether disease severity or the type of measuring device influences outcomes. Shared repositories of raw gait data, organized under consistent formats, could improve reproducibility and support more robust biomarkers for clinical applications. Finally, federated learning may also support multi-site data sharing while preserving local privacy, offering one potential approach to building scalable AI solutions [278].

7.2.2 AI Model Specificity and Technological Limitations: Chapters 3 and 4 demonstrated that LLMs can automate parts of the systematic review process, but their effectiveness and trustworthiness hinge on transparent reporting. Many studies do not disclose training data, model architectures, or validation protocols, making it difficult to assess potential biases or replicate results [279].

For HAR, classical architectures like long short-term memory (LSTM) networks have been widely used [230], and although studies sometimes omit hyperparameter details, the overall approach is generally more transparent than LLM-based screening. In contrast, LLMs suffer from limited architectural disclosure and opaque training datasets, making reproducibility and evaluation more difficult.

Adhering to guidelines and consensus statements on AI in medicine can further standardize how models are reported [280]. Studies specifically measuring recall and precision in automated screening show that high recall remains paramount for systematic reviews [274], [281]. By combining rigorous architecture disclosure, open-source practices, and privacy-preserving data exchanges, AI-based workflows could reach broader clinical acceptance without compromising security or reproducibility.

7.2.3 Data Inclusivity and Bias: Many gait and AI-oriented dementia studies predominantly address Alzheimer's disease or rely on English-language data from high-income regions, narrowing their relevance for other populations. For instance, non-AD conditions like Lewy body dementia and frontotemporal dementia can be overlooked, with fewer datasets available for rigorous analysis [282], [283]. In addition, language barriers may exclude valuable findings if research is not published in English, a gap that multilingual natural language processing methods like mBERT could help close [140].

Recent advances have also explored using AI to automate quality assessments in systematic reviews, which may help standardize evaluation and reduce variability when assessing non-English or underrepresented data sources [284]. Some clinical trials also face patterns of participant exclusion, potentially omitting older adults or specific racial/ethnic groups [285]. Expanding collaborations with low- and middle-income countries, along with broader data repositories, would create more representative clinical and observational samples. Without these measures, geographical and demographic biases can skew results and limit the applicability of findings to real-world healthcare contexts.

7.3 Future Directions

7.3.1 Refining Gait Diagnostics: Future work can explore advanced sensing technologies, such as pressure insoles or smart textiles, to capture subtle gait indicators like balance, stride symmetry, and foot

clearance. These methods could offer more precise assessments and possibly reveal early markers of neurological change. Standardizing study designs and establishing open-access repositories for gait data would also help unify sampling strategies and support reproducible research. Shared frameworks, including open-source modeling platforms [286], might further improve cross-study comparisons and strengthen the role of gait metrics in identifying different stages or subtypes of dementia.

7.3.2 Advancing Systematic Review Automation: Large language models have shown strong potential to reduce screening workloads by categorizing relevant articles more efficiently. However, many studies applying AI to systematic reviews lack technical specificity, often omitting details about their data sourcing, training protocol, and optimization strategies. Without transparency in model design and performance metrics like precision, recall, and F1-score, reproducibility and trust remain limited.

Ongoing efforts to document model architectures, validation metrics, and training data in a transparent manner would build confidence in these AI-driven workflows [274]. Developers might also align with emerging standards like CONSORT-AI and TRIPOD-AI, which provide guidelines for reporting AI in medical research [287]. Additionally, integrating explainable AI (XAI) [288] techniques such as SHAP or Grad-CAM [289] can help clarify model decision-making, fostering greater clinician confidence in automated screening tools. Open-source frameworks and shared datasets will further support reproducibility and comparison across tools.

7.3.3 Expanding Smartphone HAR to Clinical Settings: Smartphone-based activity recognition has demonstrated reliable detection of movements like walking, sitting, and standing, yet moving from pilot tests to real-world usage introduces new challenges. Domain shift can occur when participants use different phone brands, operating systems, or sensor settings, leading to inconsistent performance. Techniques like data augmentation and routine calibration may lessen these gaps [290]. Including wearable heart rate or EMG sensors can also broaden the scope of monitoring, offering more comprehensive insights into patient mobility and health.

Practical deployment in clinical environments calls for consistent documentation of sampling rates, battery demands, and hardware variations. Ensuring data privacy remains a priority; on-device processing can limit the transmission of raw sensor data, reducing security concerns. As sensor-based analytics develop further, applying sensor fusion techniques, such as Kalman filtering, to reduce measurement noise and harmonize data sources may help standardize protocols and facilitate comparisons among different research groups [290]. Federated learning also offers the option to pool insights across

multiple institutions without disclosing sensitive data [278]. These refinements may encourage broader adoption of smartphone-based assessments in routine healthcare.

7.3.4 Cross-Cutting Integration: Bridging Digital Health and Clinical Decision-Making: Combining gait diagnostics, AI-assisted reviews, and smartphone activity monitoring into routine clinical workflows will likely require cohesive data pipelines and interdisciplinary collaboration among engineering, medicine, and policy specialists. Real-time notifications could highlight emerging evidence from automated reviews or detect shifts in a patient's mobility status, prompting earlier interventions. Federated learning can coordinate these efforts at a larger scale [278], although privacy regulations may necessitate adaptive approaches to data governance. By harmonizing clinical needs with evolving technical standards, these integrated digital health tools may help deliver more responsive and person-centered care.

7.4 Concluding Remarks

This thesis has shown how computational approaches and emerging technologies can converge to address both diagnostic and translational challenges in healthcare. By analyzing gait patterns in dementia, applying large language models to systematic reviews, and developing smartphone-based human activity recognition, we have explored multiple, complementary avenues for improving patient outcomes. These methods can identify subtle mobility changes, automate parts of the research synthesis process, and provide continuous monitoring, respectively. Each domain stands to benefit from more consistent data collection, thoughtful model design, and cooperative research environments that share findings and methodologies.

Progress in these areas will require collaboration among engineers, clinicians, computer scientists, and policy specialists. Gait assessments may become more clinically relevant by adopting uniform measurement protocols and shared data repositories, while AI-assisted reviews can gain credibility through open reporting and reproducible studies. Likewise, smartphone-based tools must address sensor variability and protect personal information through privacy-aware architectures. By coordinating efforts and ensuring that ethical standards guide innovation, this multi-domain framework can help deliver proactive, person-centered care that adapts to the evolving needs of an aging and increasingly diverse population.

References

[1] V. Yazdi Feyzabadi, A. Haghdoost, M. H. Mehrolhassani, and Z. Aminian, "The Association between Peace and Life Expectancy: An Empir-ical Study of the World Countries," *Iran J Public Health*, vol. 44, no. 3, Mar. 2015, [Online]. Available: http://ijph.tums.ac.ir

[2] S. Galvani-Townsend, I. Martinez, and A. Pandey, “Is life expectancy higher in countries and territories with publicly funded health care? Global analysis of health care access and the social determinants of health,” *J Glob Health*, vol. 12, 2022, doi: 10.7189/JOGH.12.04091.

[3] J. G. B. Alves and L. V. Alves, “Early-life nutrition and adult-life outcomes,” *J Pediatr (Rio J)*, vol. 100, pp. S4–S9, Mar. 2024, doi: 10.1016/j.jped.2023.08.007.

[4] J. Cylus and L. Al Tayara, “Health, an ageing labour force, and the economy: Does health moderate the relationship between population age-structure and economic growth?,” *Soc Sci Med*, vol. 287, Oct. 2021, doi: 10.1016/j.socscimed.2021.114353.

[5] R. Osathanunkul, P. Dumrong, W. Yamaka, and P. Maneejuk, “The nonlinear impacts of aging labor and government health expenditures on productivity in ASEAN+3 economies,” *Econ Anal Policy*, vol. 80, pp. 450–470, Dec. 2023, doi: 10.1016/j.eap.2023.08.021.

[6] D. D. Le, N. T. Quashie, M. Brandt, and Y. Ibuka, “Wealth inequalities in physical and cognitive impairments across Japan and Europe: the role of health expenditure and infrastructure,” *Int J Equity Health*, vol. 22, no. 1, Dec. 2023, doi: 10.1186/s12939-023-01906-6.

[7] S. Okamoto and K. Komamura, “Towards universal health coverage in the context of population ageing: a narrative review on the implications from the long-term care system in Japan,” *Archives of Public Health*, vol. 80, no. 1, Dec. 2022, doi: 10.1186/s13690-022-00970-8.

[8] F. Longo *et al.*, “Tackling the crisis of the Italian National Health Fund,” *Lancet Public Health*, vol. 9, no. 1, pp. e6–e7, Jan. 2024, doi: 10.1016/S2468-2667(23)00277-3.

[9] R. Fenge and F. Peglow, “Decomposition of Demogra-phic Effects on the German Pension System,” *The Journal of the Economics of Ageing*, vol. 12, pp. 61–76, Nov. 2018, doi: 10.1016/j.jeoa.2018.01.001.

[10] F. Wang and J. Der Wang, “Investing preventive care and economic development in ageing societies: empirical evidences from OECD countries,” *Health Econ Rev*, vol. 11, no. 1, Dec. 2021, doi: 10.1186/s13561-021-00321-3.

[11] C. Narvaez Rojas, G. A. Alomia Peñafiel, D. F. Loaiza Buitrago, and C. A. Tavera Romero, “Society 5.0: A Japanese concept for a superintelligent society,” *Sustainability (Switzerland)*, vol. 13, no. 12, Jun. 2021, doi: 10.3390/su13126567.

[12] A. Hong, J. Welch-Stockton, J. Y. Kim, S. L. Canham, V. Greer, and M. Sorweid, “Age-Friendly Community Interventions for Health and Social Outcomes: A Scoping Review,” *Int J Environ Res Public Health*, vol. 20, no. 3, Feb. 2023, doi: 10.3390/ijerph20032554.

[13] A. Hobza and J.-T. Steeman, “Horizon Europe strategic plan 2025-2027 analysis,” Publications Office of the European Union, 2023.

[14] A. Kelley *et al.*, “National Institute on Aging’s 50th anniversary: Advancing aging research and the health and well-being of older adults,” *J Am Geriatr Soc*, vol. 72, no. 5, pp. 1574–1582, May 2024, doi: 10.1111/jgs.18837.

[15] P. Rochon *et al.*, “CIHR Institute of Aging Strategic Plan 2023-2028 Reframing Aging – Empowering Older Adults,” 2023.

[16] F. Lamontagne, K. M. Rowan, and G. Guyatt, “Integrating research into clinical practice: Challenges and solutions for Canada,” *CMAJ*, vol. 193, no. 4, pp. E127–E131, Jan. 2021, doi: 10.1503/cmaj.202397.

[17] Z. S. Morris, S. wooding, and J. Grant, “The answer is 17 years, what is the question: Understanding time lags in translational research,” *J R Soc Med*, vol. 104, no. 12, pp. 510–520, 2011, doi: 10.1258/jrsm.2011.110180.

[18] A. Garmany, S. Yamada, and A. Terzic, “Longevity leap: mind the healthspan gap,” *NPJ Regen Med*, vol. 6, no. 1, Dec. 2021, doi: 10.1038/s41536-021-00169-5.

[19] A. Garmany and A. Terzic, “Global Healthspan-Lifespan Gaps Among 183 World Health Organization Member States,” *JAMA Netw Open*, vol. 7, no. 12, p. e2450241, Dec. 2024, doi: 10.1001/jamanetworkopen.2024.50241.

[20] A. Balachandran *et al.*, “Pace of Aging in older adults matters for healthspan and lifespan.,” *medRxiv*, Apr. 2024, doi: 10.1101/2024.04.25.24306359.

[21] P. Maresova *et al.*, “Challenges and opportunity in mobility among older adults – key determinant identification,” *BMC Geriatr*, vol. 23, no. 1, Dec. 2023, doi: 10.1186/s12877-023-04106-7.

[22] T. T. Makovski, S. Schmitz, M. P. Zeegers, S. Stranges, and M. van den Akker, “Multimorbidity and quality of life: Systematic literature review and meta-analysis,” *Ageing Res Rev*, vol. 53, Aug. 2019, doi: 10.1016/j.arr.2019.04.005.

[23] P. Pedroza *et al.*, “Global and regional spending on dementia care from 2000−2019 and expected future health spending scenarios from 2020−2050: An economic modelling exercise,” *The Lancet*, vol. 45, Mar. 2022, doi: 10.1016/j.

[24] A. Brunier and A. Muchnik, “World failing to address dementia challenge,” World Health Organization. Accessed: Sep. 09, 2024. [Online]. Available: https://www.who.int/news/item/02-09-2021-world-failing-to-address-dementia-challenge

[25] B. Mikkelsen *et al.*, “Saving lives, spending less: The case for investing in noncommunicable diseases,” 2021.

[26] K. M. Thompson and D. A. Kalkowska, “An Updated Economic Analysis of the Global Polio Eradication Initiative,” *Risk Analysis*, vol. 41, no. 2, pp. 393–406, Feb. 2021, doi: 10.1111/risa.13665.

[27] F. Zhou *et al.*, "Morbidity and Mortality Weekly Report Health and Economic Benefits of Routine Childhood Immunizations in the Era of the Vaccines for Children Program-United States, 1994-2023," Aug. 2024. doi: 10.15585/mmwr.mm7331a2.

[28] S. Akter *et al.*, "Evaluation of Population-Level Tobacco Control Interventions and Health Outcomes: A Systematic Review and Meta-Analysis," *JAMA Netw Open*, vol. 6, no. 7, p. E2322341, Jul. 2023, doi: 10.1001/jamanetworkopen.2023.22341.

[29] C. M. Callahan, D. R. Bateman, S. Wang, and M. A. Boustani, "State of Science: Bridging the Science-Practice Gap in Aging, Dementia and Mental Health," *J Am Geriatr Soc*, vol. 66, pp. S28–S35, Apr. 2018, doi: 10.1111/jgs.15320.

[30] K. Bayoumy *et al.*, "Smart wearable devices in cardiovascular care: where we are and how to move forward," *Nat Rev Cardiol*, vol. 18, no. 8, pp. 581–599, Aug. 2021, doi: 10.1038/s41569-021-00522-7.

[31] M. Mason, Y. Cho, J. Rayo, Y. Gong, M. Harris, and Y. Jiang, "Technologies for Medication Adherence Monitoring and Technology Assessment Criteria: Narrative Review," *JMIR Mhealth Uhealth*, vol. 10, no. 3, Mar. 2022, doi: 10.2196/35157.

[32] J. Liutkus *et al.*, "Accuracy of a Smartphone-Based Artificial Intelligence Application for Classification of Melanomas, Melanocytic Nevi, and Seborrheic Keratoses," *Diagnostics*, vol. 13, no. 13, Jul. 2023, doi: 10.3390/diagnostics13132139.

[33] T. Kränke, K. Tripolt-Droschl, L. Röd, R. Hofmann-Wellenhof, M. Koppitz, and M. Tripolt, "New AI-algorithms on smartphones to detect skin cancer in a clinical setting—A validation study," *PLoS One*, vol. 18, no. 2 February, Feb. 2023, doi: 10.1371/journal.pone.0280670.

[34] Y. Jiang, C. Spies, J. Magin, S. J. Bhosai, L. Snyder, and J. Dunn, "Investigating the accuracy of blood oxygen saturation measurements in common consumer smartwatches," *PLOS Digital Health*, vol. 2, no. 7, Jul. 2023, doi: 10.1371/journal.pdig.0000296.

[35] J. Rafl, T. E. Bachman, V. Rafl-Huttova, S. Walzel, and M. Rozanek, "Commercial smartwatch with pulse oximeter detects short-time hypoxemia as well as standard medical-grade device: Validation study," *Digit Health*, vol. 8, 2022, doi: 10.1177/20552076221132127.

[36] J. Wilson, L. Allcock, R. Mc Ardle, J.-P. Taylor, and L. Rochester, "The neural correlates of discrete gait characteristics in ageing- A structured review," *Neurosci Biobehav Rev*, vol. 100, pp. 344–369, May 2019, doi: 10.1016/j.neubiorev.2018.12.017.

[37] E. Buckley, C. Mazzà, and A. McNeill, "A systematic review of the gait characteristics associated with Cerebellar Ataxia," *Gait Posture*, vol. 60, pp. 154–163, Dec. 2018, doi: 10.1016/j.gaitpost.2017.11.024.

[38] S. Fritz and M. Lusardi, "White Paper: 'Walking Speed: the Sixth Vital Sign,'" *Journal of Geriatric Physical Therapy*, vol. 32, no. 2, pp. 2–5, 2009, doi: 10.1519/00139143-200932020-00002.

[39] B. Jin *et al.*, "Walking-age analyzer for healthcare applications," *IEEE J Biomed Health Inform*, vol. 18, no. 3, pp. 1034–1042, 2014, doi: 10.1109/JBHI.2013.2296873.

[40] A. I. Faisal, S. Majumder, T. Mondal, D. Cowan, S. Naseh, and M. J. Deen, "Monitoring methods of human body joints: State-of-the-art and research challenges," *Sensors (Switzerland)*, vol. 19, no. 11, Jun. 2019, doi: 10.3390/s19112629.

[41] P. Mandal, K. Tank, T. Mondal, C. H. Chen, and M. J. Deen, "Predictive Walking-Age Health Analyzer," *IEEE J Biomed Health Inform*, vol. 22, no. 2, pp. 363–374, Mar. 2018, doi: 10.1109/JBHI.2017.2666603.

[42] A. I. Faisal, T. Mondal, D. Cowan, and M. J. Deen, "Characterization of Knee and Gait Features from a Wearable Tele-Health Monitoring System," *IEEE Sens J*, vol. 22, no. 6, pp. 4741–4753, Mar. 2022, doi: 10.1109/JSEN.2022.3146617.

[43] R. Baker, "The history of gait analysis before the advent of modern computers," *Gait Posture*, vol. 26, no. 3, pp. 331–342, Sep. 2007, doi: 10.1016/j.gaitpost.2006.10.014.

[44] D. Cattaneo *et al.*, "Mobility disorders in Stroke, Parkinson disease, and multiple sclerosis: A multicenter cross-sectional study," *Am J Phys Med Rehabil*, vol. 99, no. 1, pp. 41–47, Jan. 2020, doi: 10.1097/PHM.0000000000001272.

[45] H. B. Menz, M. D. Latt, A. Tiedemann, M. M. S. Kwan, and S. R. Lord, "Reliability of the GAITRite® walkway system for the quantification of temporo-spatial parameters of gait in young and older people," *Gait Posture*, vol. 20, no. 1, pp. 20–25, Aug. 2004, doi: 10.1016/S0966-6362(03)00068-7.

[46] D. Vun Yee Sing, R. Bowers, and A. McGarry, "Vision-based motion capture for the gait analysis of neurodegenerative diseases- A review ," *Gait Posture*, vol. 112, pp. 95–107, Jul. 2024, doi: 10.1016/j.gaitpost.2024.04.029.

[47] E. Y. Boateng, J. Otoo, and D. A. Abaye, "Basic Tenets of Classification Algorithms K-Nearest-Neighbor, Support Vector Machine, Random Forest and Neural Network: A Review," *Journal of Data Analysis and Information Processing*, vol. 08, no. 04, pp. 341–357, 2020, doi: 10.4236/jdaip.2020.84020.

[48] G. Dion *et al.*, "In-sensor human gait analysis with machine learning in a wearable microfabricated accelerometer," *Communications Engineering*, vol. 3, no. 1, Dec. 2024, doi: 10.1038/s44172-024-00193-5.

[49] X. Liu *et al.*, “Wearable Devices for Gait Analysis in Intelligent Healthcare,” May 13, 2021, *Frontiers Media S.A.* doi: 10.3389/fcomp.2021.661676.

[50] H. Prasanth *et al.*, “Wearable sensor-based real-time gait detection: A systematic review,” *Sensors*, vol. 21, no. 8, Apr. 2021, doi: 10.3390/s21082727.

[51] V. C. Pereira, S. N. Silva, V. K. S. Carvalho, F. Zanghelini, and J. O. M. Barreto, “Strategies for the implementation of clinical practice guidelines in public health: an overview of systematic reviews,” *Health Res Policy Syst*, vol. 20, no. 1, Dec. 2022, doi: 10.1186/s12961-022-00815-4.

[52] H. Khalil *et al.*, “Conducting high quality scoping reviews-challenges and solutions,” *J Clin Epidemiol*, vol. 130, pp. 156–160, Feb. 2021, doi: 10.1016/j.jclinepi.2020.10.009.

[53] Z. Munn, C. Stern, E. Aromataris, C. Lockwood, and Z. Jordan, “What kind of systematic review should i conduct? A proposed typology and guidance for systematic reviewers in the medical and health sciences,” *BMC Med Res Methodol*, vol. 18, no. 1, Jan. 2018, doi: 10.1186/s12874-017-0468-4.

[54] A. C. Tricco, S. E. Straus, A. Ghaffar, and E. V. Langlois, “Rapid reviews for health policy and systems decision-making: more important than ever before,” *Syst Rev*, vol. 11, no. 1, Dec. 2022, doi: 10.1186/s13643-022-01887-7.

[55] T. Lasserson *et al.*, “Standards for the reporting of protocols of new Cochrane Intervention reviews,” *Methodological expectations of Cochrane intervention reviews*, Feb. 2022, Accessed: Oct. 07, 2022. [Online]. Available: https://community.cochrane.org/mecir-manual

[56] R. Sarkis-Onofre, F. Catalá-López, E. Aromataris, and C. Lockwood, “How to properly use the PRISMA Statement,” *Syst Rev*, vol. 10, no. 1, Dec. 2021, doi: 10.1186/s13643-021-01671-z.

[57] D. Moher *et al.*, “Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement,” *Revista Espanola de Nutricion Humana y Dietetica*, vol. 20, no. 2, pp. 148–160, 2016, doi: 10.1186/2046-4053-4-1.

[58] H. Khalil *et al.*, “Advancing the methodology of mapping reviews: A scoping review,” *Res Synth Methods*, vol. 15, no. 3, pp. 384–397, May 2024, doi: 10.1002/jrsm.1694.

[59] D. F. Stroup *et al.*, “Meta-analysis of Observational Studies in Epidemiology: A Proposal for Reporting,” *JAMA*, vol. 283, no. 15, p. 2008, 2000, doi: 10.1001/jama.283.15.2008.

[60] L. Kellermeyer, B. Harnke, and S. Knight, “Covidence and Rayyan,” *Journal of the Medical Library Association*, vol. 106, no. 4, pp. 580–583, Oct. 2018, doi: 10.5195/JMLA.2018.513.

[61] “Review Manager (RevMan),” 2020, *The Nordic Cochrane Centre, Copenhagen*: 5.4. Accessed: Jan. 20, 2022. [Online]. Available: revman.cochrane.org

[62] K. Bullers *et al.*, "It takes longer than you think: Librarian time spent on systematic review tasks," *Journal of the Medical Library Association*, vol. 106, no. 2, pp. 198–207, Apr. 2018, doi: 10.5195/jmla.2018.323.

[63] R. Borah, A. W. Brown, P. L. Capers, and K. A. Kaiser, "Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry," *Open*, vol. 7, p. 12545, 2017, doi: 10.1136/bmjopen-2016.

[64] L. Schmidt, B. K. Olorisade, L. A. McGuinness, and J. P. T. Higgins, "Data extraction methods for systematic review (semi)automation: A living review protocol," *F1000Res*, vol. 9, p. 210, Mar. 2020, doi: 10.12688/f1000research.22781.1.

[65] A. O'Mara-Eves, J. Thomas, J. McNaught, M. Miwa, and S. Ananiadou, "Using text mining for study identification in systematic reviews: A systematic review of current approaches," *Syst Rev*, vol. 4, no. 1, Jan. 2015, doi: 10.1186/2046-4053-4-5.

[66] R. Ofori-Boateng, M. Aceves-Martins, N. Wiratunga, and C. F. Moreno-Garcia, "Towards the automation of systematic reviews using natural language processing, machine learning, and deep learning: a comprehensive review," *Artif Intell Rev*, vol. 57, no. 8, Aug. 2024, doi: 10.1007/s10462-024-10844-w.

[67] F. Dennstädt, J. Zink, P. M. Putora, J. Hastings, and N. Cihoric, "Title and abstract screening for literature reviews using large language models: an exploratory study in the biomedical domain," *Syst Rev*, vol. 13, no. 1, Dec. 2024, doi: 10.1186/s13643-024-02575-4.

[68] K. B. Rajan, J. Weuve, L. L. Barnes, E. A. McAninch, R. S. Wilson, and D. A. Evans, "Population estimate of people with clinical Alzheimer's disease and mild cognitive impairment in the United States (2020–2060)," *Alzheimer's and Dementia*, vol. 17, no. 12, pp. 1966–1975, Dec. 2021, doi: 10.1002/alz.12362.

[69] M. Fang *et al.*, "Lifetime risk and projected burden of dementia," *Nat Med*, Jan. 2025, doi: 10.1038/s41591-024-03340-9.

[70] D. Morgan, M. Funk, M. Crossley, J. Basran, A. Kirk, and V. Dal Bello-Haas, "The Potential of Gait Analysis to Contribute to Differential Diagnosis of Early Stage Dementia: Current Research and Future Directions," *Canadian Journal on Aging*, vol. 26, no. 1, 2007, doi: 10.3138/1457-2411-v402-62l1.

[71] Z. Arvanitakis, R. C. Shah, and D. A. Bennett, "Diagnosis and Management of Dementia: Review," *JAMA - Journal of the American Medical Association*, vol. 322, no. 16, pp. 1589–1599, Oct. 2019, doi: 10.1001/jama.2019.4782.

[72] S. Duong, T. Patel, and F. Chang, "Dementia: What pharmacists need to know," *Canadian Pharmacists Journal*, vol. 150, no. 2, pp. 118–129, Mar. 2017, doi: 10.1177/1715163517690745.

[73] C. U. Onyike and J. Diehl-Schmid, "The epidemiology of frontotemporal dementia," *International Review of Psychiatry*, vol. 25, no. 2, pp. 130–137, Apr. 2013, doi: 10.3109/09540261.2013.776523.

[74] M. F. Weiner *et al.*, "Can Alzheimer's Disease and Dementias with Lewy Bodies Be Distinguished Clinically?," *J Geriatr Psychiatry Neurol*, vol. 16, no. 4, pp. 245–250, Dec. 2003, doi: 10.1177/0891988703258671.

[75] R. Mc Ardle, B. Galna, P. Donaghy, A. Thomas, and L. Rochester, "Do Alzheimer's and Lewy body disease have discrete pathological signatures of gait?," *Alzheimer's and Dementia*, vol. 15, no. 10, pp. 1367–1377, Oct. 2019, doi: 10.1016/j.jalz.2019.06.4953.

[76] M. B. Van Iersel, W. Hoefsloot, M. Munneke, B. R. Bloem, and M. G. M. Olde Rikkert, "Systematic review of quantitative clinical gait analysis in patients with dementia," *Z Gerontol Geriatr*, vol. 37, no. 1, pp. 27–32, Feb. 2004, doi: 10.1007/s00391-004-0176-7.

[77] R. Chiaramonte and M. Cioni, "Critical spatiotemporal gait parameters for individuals with dementia: A systematic review and meta-analysis," *Hong Kong Physiotherapy Journal*, vol. 41, no. 1, pp. 1–14, Jun. 2021, doi: 10.1142/S101370252130001X.

[78] M. J. Deen, "Information and communications technologies for elderly ubiquitous healthcare in a smart home," *Pers Ubiquitous Comput*, vol. 19, no. 3–4, pp. 573–599, Jul. 2015, doi: 10.1007/s00779-015-0856-x.

[79] J. Cohen, *Statistical power analysis for the behavioral sciences*, 2nd ed. New York: Lawrence Erlbaum Associates, 2013. doi: 10.4324/9780203771587.

[80] Wells GA, Shea B, O'Connell D, Peterson J, Welch V, and Losos M, "The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomized studies in meta-analyses," Ottawa: Ottawa Hospital Research Institute. Accessed: May 15, 2023. [Online]. Available: https://www.ohri.ca/programs/clinical_epidemiology/oxford.asp

[81] J. E. Wittwer, K. E. Webster, and H. B. Menz, "A longitudinal study of measures of walking in people with Alzheimer's Disease," *Gait Posture*, vol. 32, no. 1, pp. 113–117, May 2010, doi: 10.1016/j.gaitpost.2010.04.001.

[82] J. R. Merory, J. E. Wittwer, C. C. Rowe, and K. E. Webster, "Quantitative gait analysis in patients with dementia with Lewy bodies and Alzheimer's disease," *Gait Posture*, vol. 26, no. 3, pp. 414–419, Sep. 2007, doi: 10.1016/j.gaitpost.2006.10.006.

[83] R. Mc Ardle, S. Del Din, P. Donaghy, B. Galna, A. J. Thomas, and L. Rochester, "The impact of environment on gait assessment: Considerations from real-world gait analysis in dementia subtypes," *Sensors (Switzerland)*, vol. 21, no. 3, pp. 1–15, Feb. 2021, doi: 10.3390/s21030813.

[84] L. Ni *et al.*, "Pathological Gait Signatures of Post-stroke Dementia With Toe-Off and Heel-to-Ground Angles Discriminate From Alzheimer's Disease," *Front Aging Neurosci*, vol. 13, Nov. 2021, doi: 10.3389/fnagi.2021.766884.

[85] S. G. Velayutham, S. R. Chandra, S. Bharath, and R. G. Shankar, "Quantitative balance and gait measurement in patients with frontotemporal dementia and Alzheimer diseases: A pilot study," *Indian J Psychol Med*, vol. 39, no. 2, pp. 176–182, Mar. 2017, doi: 10.4103/0253-7176.203132.

[86] G. Allali *et al.*, "Frontotemporal dementia: Pathology of gait?," *Movement Disorders*, vol. 25, no. 6, pp. 731–737, Apr. 2010, doi: 10.1002/mds.22927.

[87] P.-C. Chung, Y.-L. Hsu, C.-Y. Wang, C.-W. Lin, J.-S. Wang, and M.-C. Pai, "Gait analysis for patients with Alzheimer'S disease using a triaxial accelerometer," *2012 IEEE International Symposium on Circuits and Systems*, pp. 1323–1326, May 2012, doi: 10.1109/ISCAS.2012.6271484.

[88] Y. L. Hsu *et al.*, "Gait and balance analysis for patients with Alzheimer's disease using an inertial-sensor-based wearable instrument," *IEEE J Biomed Health Inform*, vol. 18, no. 6, pp. 1822–1830, Nov. 2014, doi: 10.1109/JBHI.2014.2325413.

[89] Y. C. Lin, W. C. Hsu, C. K. Wu, W. H. Chang, K. P. H. Wu, and A. M. K. Wong, "Comparison of motor performance of upper and lower extremities in dual-task tests in patients with mild Alzheimer's dementia," *Aging Clin Exp Res*, vol. 28, no. 3, pp. 491–496, Jun. 2016, doi: 10.1007/s40520-015-0441-1.

[90] Xiaoli Meng, Haoyong Yu, and Ming Po Tham, "Gait phase detection in able-bodied subjects and dementia patients," *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 4907–4910, Jul. 2013, doi: 10.1109/EMBC.2013.6610648.

[91] N. K. Nadkarni, E. Mawji, W. E. McIlroy, and S. E. Black, "Spatial and temporal gait parameters in Alzheimer's disease and aging," *Gait Posture*, vol. 30, no. 4, pp. 452–454, Nov. 2009, doi: 10.1016/j.gaitpost.2009.07.003.

[92] N. K. Nadkarni, W. E. McIlroy, E. Mawji, and S. E. Black, "Gait and subcortical hyperintensities in mild alzheimer's disease and aging," *Dement Geriatr Cogn Disord*, vol. 28, no. 4, pp. 295–301, Nov. 2009, doi: 10.1159/000245158.

[93] N. Ogama, H. Endo, S. Satake, S. Niida, H. Arai, and T. Sakurai, "Impact of regional white matter hyperintensities on specific gait function in Alzheimer's disease and mild cognitive impairment," *J Cachexia Sarcopenia Muscle*, vol. 12, no. 6, pp. 2045–2055, Dec. 2021, doi: 10.1002/jcsm.12807.

[94] A. Tanaka, H. Okuzumi, I. Kobayashi, N. Murai, K. Meguro, and T. N. Akamura, "Gait disturbance of patients with vascular and Alzheimer-type dementias," *Percept Mot Skills*, vol. 80, no. 3, pp. 735–738, Jun. 1995, doi: 10.2466/pms.1995.80.3.735.

[95] Q. Duan *et al.*, "Gait Domains May Be Used as an Auxiliary Diagnostic Index for Alzheimer's Disease," *Brain Sci*, vol. 13, no. 11, p. 1599, Nov. 2023, doi: 10.3390/brainsci13111599.

[96] R. Ma, H. Zhào, W. Wei, Y. Liu, and Y. Huang, "Gait characteristics under single-/dual-task walking conditions in elderly patients with cerebral small vessel disease: Analysis of gait variability, gait asymmetry and bilateral coordination of gait," *Gait Posture*, vol. 92, pp. 65–70, Feb. 2022, doi: 10.1016/j.gaitpost.2021.11.007.

[97] Y. Hou, Y. Li, S. Yang, W. Qin, L. Yang, and W. Hu, "Gait Impairment and Upper Extremity Disturbance Are Associated With Total Magnetic Resonance Imaging Cerebral Small Vessel Disease Burden," *Front Aging Neurosci*, vol. 13, May 2021, doi: 10.3389/fnagi.2021.640844.

[98] G. Allali *et al.*, "Gait phenotype from mild cognitive impairment to moderate dementia: Results from the GOOD initiative," *Eur J Neurol*, vol. 23, no. 3, pp. 527–541, Mar. 2016, doi: 10.1111/ene.12882.

[99] L. Simieli, F. A. Barbieri, D. Orcioli-Silva, E. Lirani-Silva, F. Stella, and L. T. B. Gobbi, "Obstacle crossing with dual tasking is a danger for individuals with Alzheimer's disease and for healthy older people," *Journal of Alzheimer's Disease*, vol. 43, no. 2, pp. 435–441, 2015, doi: 10.3233/JAD-140807.

[100] C. Oh, "Single-Task or Dual-Task? Gait Assessment as a Potential Diagnostic Tool for Alzheimer's Dementia," *Journal of Alzheimer's Disease*, vol. 84, no. 3, pp. 1183–1192, Nov. 2021, doi: 10.3233/JAD-210690.

[101] F. A. Barbieri, L. Simieli, D. Orcioli-Silva, R. Vito´rio, F. Stella, and L. T. Bucken Gobbi, "Variability in Obstacle Clearance May (Not) Indicate Cognitive Disorders in Alzheimer Disease," *Alzheimer Dis Assoc Disord*, vol. 20, no. 4, pp. 307–311, 2014, doi: 10.1097/WAD.0000000000000063.

[102] D. A. Kegelmeyer *et al.*, "Immediate effects of treadmill walking in individuals with Lewy body dementia and Huntington's disease," *Gait Posture*, vol. 86, pp. 186–191, May 2021, doi: 10.1016/j.gaitpost.2021.03.016.

[103] N. B. Alexander *et al.*, "Maintenance of balance, gait patterns, and obstacle clearance in Alzheimer's disease," *Neurology*, vol. 45, no. 5, pp. 908–914, 1995, doi: 10.1212/wnl.45.5.908.

[104] J. H. Ansai, T. H. Nakagawa, L. P. De Andrade, P. Giusti Rossi, F. A. C. Vale, and J. R. Rebelatto, "Differences in timed up and go subtasks between older people with mild cognitive

impairment and mild Alzheimer's disease," *Motor Control*, vol. 23, no. 1, pp. 1–12, Jan. 2019, doi: 10.1123/mc.2017-0015.

[105] S. Gillain *et al.*, "The value of instrumental gait analysis in elderly healthy, MCI or Alzheimer's disease subjects and a comparison with other clinical tests used in single and dual-task conditions," *Ann Phys Rehabil Med*, vol. 52, no. 6, pp. 453–474, Jul. 2009, doi: 10.1016/j.rehab.2008.10.004.

[106] W. Goldman, J. Baty, V. Buckles, S. Sahrmann, and J. Morris, "Motor dysfunction in mildly demented AD individuals without extrapyramidal signs," *Neurology*, vol. 53, no. 5, pp. 956–956, 1998, doi: 10.1212/wnl.53.5.956.

[107] L. Z. Gras, S. F. Kanaan, J. M. McDowd, Y. M. Colgrove, J. Burns, and P. S. Pohl, "Balance and gait of adults with very mild Alzheimer disease," *Journal of Geriatric Physical Therapy*, vol. 38, no. 1, pp. 1–7, Jan. 2015, doi: 10.1519/JPT.0000000000000020.

[108] S. W. Hunter *et al.*, "Effect of Learning to Use a Mobility Aid on Gait and Cognitive Demands in People with Mild to Moderate Alzheimer's Disease: Part i-Cane," *Journal of Alzheimer's Disease*, vol. 71, no. s1, pp. S105–S114, 2019, doi: 10.3233/JAD-181169.

[109] D. Maquet *et al.*, "Gait analysis in elderly adult patients with mild cognitive impairment and patients with mild Alzheimer's disease: Simple versus dual task: A preliminary report," *Clin Physiol Funct Imaging*, vol. 30, no. 1, pp. 51–56, Jan. 2010, doi: 10.1111/j.1475-097X.2009.00903.x.

[110] S. W. Muir, M. Speechley, J. Wells, M. Borrie, K. Gopaul, and M. Montero-Odasso, "Gait assessment in mild cognitive impairment and Alzheimer's disease: The effect of dual-task challenges across the cognitive spectrum," *Gait Posture*, vol. 35, no. 1, pp. 96–100, Jan. 2012, doi: 10.1016/j.gaitpost.2011.08.014.

[111] F. Pieruccini-Faria *et al.*, "Gait variability across neurodegenerative and cognitive disorders: Results from the Canadian Consortium of Neurodegeneration in Aging (CCNA) and the Gait and Brain Study," *Alzheimer's and Dementia*, vol. 17, no. 8, pp. 1317–1328, Aug. 2021, doi: 10.1002/alz.12298.

[112] P. Suttanon, K. D. Hill, C. M. Said, D. LoGiudice, N. T. Lautenschlager, and K. J. Dodd, "Balance and mobility dysfunction and falls risk in older people with mild to moderate alzheimer disease," *Am J Phys Med Rehabil*, vol. 91, no. 1, pp. 12–23, Feb. 2012, doi: 10.1097/PHM.0b013e31823caeea.

[113] J. Y. Tung *et al.*, "Measuring life space in older adults with mild-to-moderate Alzheimer's disease using mobile phone GPS," *Gerontology*, vol. 60, no. 2, pp. 154–162, 2014, doi: 10.1159/000355669.

[114] H. Visser, “Gait and balance in senile dementia of Alzheimer’s type,” *Age Ageing*, vol. 12, no. 4, pp. 296–301, 1983, doi: 10.1093/ageing/12.4.296.

[115] F. A. A. Masse, J. H. Ansai, R. C. Gerassi, C. Tsen, N. O. de C. Cezar, and L. P. de Andrade, “Six-month change in gait speed to discriminate between those with and without falls history in older people with Mild Cognitive Impairment and mild Alzheimer disease,” *Geriatr Nurs (Minneap)*, vol. 48, pp. 274–279, Nov. 2022, doi: 10.1016/j.gerinurse.2022.10.002.

[116] A. Sainsily-Cesarus *et al.*, “Dementia with Lewy bodies and gait neural basis: a cross-sectional study,” *Alzheimers Res Ther*, vol. 16, no. 1, p. 170, Jul. 2024, doi: 10.1186/s13195-024-01539-z.

[117] X. Weng *et al.*, “White matter hyperintensities: a possible link between sarcopenia and cognitive impairment in patients with mild to moderate Alzheimer’s disease,” *Eur Geriatr Med*, vol. 14, no. 5, pp. 1037–1047, Jun. 2023, doi: 10.1007/s41999-023-00818-6.

[118] J. Plácido *et al.*, “Beyond the Mini-Mental State Examination: The Use of Physical and Spatial Navigation Tests to Help to Screen for Mild Cognitive Impairment and Alzheimer’s Disease,” *Journal of Alzheimer’s Disease*, vol. 81, no. 3, pp. 1243–1252, Jun. 2021, doi: 10.3233/JAD-210106.

[119] R. Mc Ardle, S. Del Din, B. Galna, A. Thomas, and L. Rochester, “Differentiating dementia disease subtypes with gait analysis: feasibility of wearable sensors?,” *Gait Posture*, vol. 76, pp. 372–376, Feb. 2020, doi: 10.1016/j.gaitpost.2019.12.028.

[120] K. E. Webster, J. R. Merory, and J. E. Wittwer, “Gait Variability in Community Dwelling Adults With Alzheimer Disease,” *Alzheimer Dis Assoc Disord*, vol. 20, no. 1, pp. 37–40, 2008, doi: 10.1097/01.wad.0000201849.75578.de.

[121] J. E. Wittwer, P. T. Andrews, K. E. Webster, and H. B. Menz, “Timing variability during gait initiation is increased in people with Alzheimer’s disease compared to controls,” *Dement Geriatr Cogn Disord*, vol. 26, no. 3, pp. 277–283, Oct. 2008, doi: 10.1159/000160961.

[122] J. S. Choi *et al.*, “Comparison of gait and cognitive function among the elderly with alzheimer’s disease, mild cognitive impairment and healthy,” *International Journal of Precision Engineering and Manufacturing*, vol. 12, no. 1, pp. 169–173, Feb. 2011, doi: 10.1007/s12541-011-0024-9.

[123] N. E. Fritz *et al.*, “Motor performance differentiates individuals with Lewy body dementia, Parkinson’s and Alzheimer’s disease,” *Gait Posture*, vol. 50, pp. 1–7, Oct. 2016, doi: 10.1016/j.gaitpost.2016.08.009.

[124] K. K. Gnanalingham, J. Byrne, A. Thornton, M. A. Sambrook, and P. Bannister, “Motor and cognitive function in Lewy body dementia: comparison with Alzheimer’s and Parkinson’s diseases,” 1997. doi: 10.1136/jnnp.62.3.243.

[125] G. D. Rabinovici *et al.*, "Distinct MRI atrophy patterns in autopsy-proven Alzheimer's disease and frontotemporal lobar degeneration," *Am J Alzheimers Dis Other Demen*, vol. 22, no. 6, pp. 474–488, 2008, doi: 10.1177/1533317507308779.

[126] G. Allali *et al.*, "Brain structure covariance associated with gait control in aging," *Journals of Gerontology - Series A Biological Sciences and Medical Sciences*, vol. 74, no. 5, pp. 705–713, May 2019, doi: 10.1093/gerona/gly123.

[127] L. M. Allan, C. G. Ballard, D. J. Burn, and R. A. Kenny, "Prevalence and severity of gait disorders in Alzheimer's and non-Alzheimer's dementias," *J Am Geriatr Soc*, vol. 53, no. 10, pp. 1681–1687, Oct. 2005, doi: 10.1111/j.1532-5415.2005.53552.x.

[128] D. Gough and D. Elbourne, "Systematic Research Synthesis to Inform Policy, Practice and Democratic Debate," *Social Policy and Society*, vol. 1, no. 3, pp. 225–236, Jul. 2002, doi: 10.1017/s147474640200307x.

[129] M. Michelson and K. Reuter, "The significant cost of systematic reviews and meta-analyses: A call for greater involvement of machine learning to assess the promise of clinical trials," *Contemp Clin Trials Commun*, vol. 16, Dec. 2019, doi: 10.1016/j.conctc.2019.100443.

[130] R. Qureshi, D. Shaughnessy, K. A. R. Gill, K. A. Robinson, T. Li, and E. Agai, "Are ChatGPT and large language models 'the answer' to bringing us closer to systematic review automation?," *Syst Rev*, vol. 12, no. 1, Dec. 2023, doi: 10.1186/s13643-023-02243-z.

[131] P. Blumenfeld *et al.*, "The lag time in initiating clinical testing of new drugs in combination with radiation therapy, a significant barrier to progress?," *Br J Cancer*, vol. 111, no. 7, pp. 1305–1309, Sep. 2014, doi: 10.1038/bjc.2014.448.

[132] G. A. Van Norman, "Translational Toolbox Drugs, Devices, and the FDA: Part 2 An Overview of Approval Processes: FDA Approval of Medical Devices," *Basic Trans Science*, vol. 1, no. 1, pp. 277–87, 2016, Accessed: Jan. 24, 2024. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/30167516/

[133] C. Lefebvre *et al.*, "Searching for and selecting studies," in *Cochrane Handbook for Systematic Reviews of Interventions*, Wiley, 2019, pp. 67–107. doi: 10.1002/9781119536604.ch4.

[134] B. C. Wallace, T. A. Trikalinos, J. Lau, C. Brodley, and C. H. Schmid, "Semi-automated screening of biomedical citations for systematic reviews," *BMC Bioinformatics*, vol. 11, no. 55, 2010, doi: 10.1186/1471-2105-11-55.

[135] Institute of Medicine *et al.*, *Finding What Works in Health Care*. Washington, D.C.: National Academies Press, 2011. doi: 10.17226/13059.

[136] A. H. Razavi, S. Matwin, D. Inkpen, and A. Kouznetsov, "Parameterized contrast in second order soft co-occurrences: A novel text representation technique in text mining and knowledge

extraction," in *ICDM Workshops 2009 - IEEE International Conference on Data Mining*, 2009, pp. 471–476. doi: 10.1109/ICDMW.2009.49.

[137] B. C. Wallace, K. Small, C. E. Brodley, J. Lau, and T. A. Trikalinos, "Modeling annotation time to reduce workload in comparative effectiveness reviews," in *Proceedings of the 1st ACM International Health Informatics Symposium*, New York, NY, USA: ACM, Nov. 2010, pp. 28–35. doi: 10.1145/1882992.1882999.

[138] D. Reichenpfader, H. Müller, and K. Denecke, "Large language model-based information extraction from free-text radiology reports: a scoping review protocol," *BMJ Open*, vol. 13, no. 12, Dec. 2023, doi: 10.1136/bmjopen-2023-076865.

[139] T. B. Brown *et al.*, "Language Models are Few-Shot Learners," *ArXiv Computation and Language*, 2020, doi: 10.48550/arXiv.2005.14165.

[140] J. Devlin, M.-W. Chang, K. Lee, K. T. Google, and A. I. Language, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," 2019. [Online]. Available: https://github.com/tensorflow/tensor2tensor

[141] A. Birhane, A. Kasirzadeh, D. Leslie, and S. Wachter, "Science in the age of large language models," *Nature Reviews Physics*, vol. 5, no. 5, pp. 277–280, May 2023, doi: 10.1038/s42254-023-00581-4.

[142] A. Vaswani *et al.*, "Attention Is All You Need," *ArXiv Computation and Language*, 2017, doi: 10.48550/arXiv.1706.03762.

[143] OpenAI *et al.*, "GPT-4 Technical Report," *ArXiv Computation and Language*, Mar. 2023, Accessed: Mar. 14, 2023. [Online]. Available: http://arxiv.org/abs/2303.08774

[144] R. Anil *et al.*, "PaLM 2 Technical Report," *ArXiv Computation and Language*, May 2023, Accessed: Jan. 25, 2024. [Online]. Available: http://arxiv.org/abs/2305.10403

[145] M. Shoeybi, M. Patwary, R. Puri, P. Legresley, J. Casper, and B. Catanzaro, "Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism," *ArXiv Computation and Language*, 2020, doi: 10.48550/arXiv.1909.08053.

[146] M. A. Hearst, "Untangling Text Data Mining," *Proceedings of the 37th Annual meeting of the Association for Computational Linguistics*, pp. 3–10, 1999, doi: 10.3115/1034678.1034679.

[147] Ö. Aydin and E. Karaarslan, "OpenAI ChatGPT Generated Literature Review: Digital Twin in Healthcare," *Emerging Computer Technologies*, vol. 2, pp. 22–31, 2022, doi: 10.2139/ssrn.4308687.

[148] M. Cascella, J. Montomoli, V. Bellini, and E. Bignami, "Evaluating the Feasibility of ChatGPT in Healthcare: An Analysis of Multiple Clinical and Research Scenarios," *J Med Syst*, vol. 47, no. 1, Dec. 2023, doi: 10.1007/s10916-023-01925-4.

[149] R. Vaishya, A. Misra, and A. Vaish, "ChatGPT: Is this version good for healthcare and research?," *Diabetes and Metabolic Syndrome: Clinical Research and Reviews*, vol. 17, no. 4, Apr. 2023, doi: 10.1016/j.dsx.2023.102744.

[150] M. Halaweh, "ChatGPT in education: Strategies for responsible implementation," *Contemp Educ Technol*, vol. 15, no. 2, Apr. 2023, doi: 10.30935/cedtech/13036.

[151] T. H. Kung *et al.*, "Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models," *PLOS Digital Health*, vol. 2, no. 2, p. e0000198, Feb. 2023, doi: 10.1371/journal.pdig.0000198.

[152] X. Zhai, "ChatGPT for Next Generation Science Learning," *XRDS: Crossroads, The ACM Magazine for Students*, vol. 29, no. 3, pp. 42–46, 2023, doi: 10.2139/ssrn.4331313.

[153] J. Rudolph, S. Tan, and S. Tan, "ChatGPT: Bullshit spewer or the end of traditional assessments in higher education?," *Journal of Applied Learning and Teaching*, vol. 6, no. 1, pp. 342–363, Jan. 2023, doi: 10.37074/jalt.2023.6.1.9.

[154] S. A. Prieto, E. T. Mengiste, and B. García de Soto, "Investigating the Use of ChatGPT for the Scheduling of Construction Projects," *Buildings*, vol. 13, no. 4, Apr. 2023, doi: 10.3390/buildings13040857.

[155] H. You, Y. Ye, T. Zhou, Q. Zhu, and J. Du, "Robot-Enabled Construction Assembly with Automated Sequence Planning based on ChatGPT: RoboGPT," 2023, doi: 10.1109/ACCESS.2022.Doi.

[156] H. Alkaissi and S. I. McFarlane, "Artificial Hallucinations in ChatGPT: Implications in Scientific Writing," *Cureus*, Feb. 2023, doi: 10.7759/cureus.35179.

[157] M. Salvagno, F. S. Taccone, and A. G. Gerli, "Can artificial intelligence help for scientific writing?," *Crit Care*, vol. 27, no. 1, Dec. 2023, doi: 10.1186/s13054-023-04380-2.

[158] H. Zheng and H. Zhan, "ChatGPT in Scientific Writing: A Cautionary Tale," *American Journal of Medicine*, vol. 136, no. 8, pp. 725-726.e6, Aug. 2023, doi: 10.1016/j.amjmed.2023.02.011.

[159] R. Luo, S. Subramaniam, A. Ilius Faisal, L. Mbuagbaw, and M. Jamal Deen, "Gait spatiotemporal parameters in dementia subtypes: systematic review and meta-analysis," *Manuscript submitted for publication, 2025*, 2025.

[160] Z. Sastimoglu, S. Subramaniam, A. I. Faisal, W. Jiang, A. Ye, and M. J. Deen, "Wearable PPG Based BP Estimation Methods: A Systematic Review and Meta-Analysis," *Journal of Biomedical Health Informatics*, vol. 29, no. 4, pp. 2439–2452, Apr. 2025, doi: 10.1109/JBHI.2024.3499834.

[161] A. Ilius, W. Jiang, S. Subramaniam, L. Mbuagbaw, and M. J. Deen, "Impact of Age and Comorbidities on Long COVID: A Systematic Review and Meta-Analysis," *Manuscript submitted for publication, 2025*, 2025.

[162] Open Source, "LangChain: Build context-aware reasoning applications," 2022, *GitHub*: 0.0.27. Accessed: May 29, 2024. [Online]. Available: https://js.langchain.com/docs/introduction/

[163] J. T. Minion *et al.*, "PICO Portal," *Journal of the Canadian Health Libraries Association*, vol. 42, no. 3, pp. 181–183, Dec. 2021, doi: 10.29173/JCHLA29590.

[164] V. Taieb, B. Smela-Lipińska, P. O'Blenis, and C. François, "PRM181 - Use Of Artificial Intelligence With Distillersr Software For A Systematic Literature Review Of Utilities In Infectious Disease," *Value in Health*, vol. 21, p. S387, Oct. 2018, doi: 10.1016/j.jval.2018.09.2299.

[165] I. J. Marshall, J. Kuiper, E. Banner, and B. C. Wallace, "Automating biomedical evidence synthesis: Robotreviewer," in *ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of System Demonstrations*, Association for Computational Linguistics (ACL), 2017, pp. 7–12. doi: 10.18653/v1/P17-4002.

[166] I. J. Marshall *et al.*, "Trialstreamer: A living, automatically updated database of clinical trial reports," *Journal of the American Medical Informatics Association*, vol. 27, no. 12, pp. 1903–1912, Dec. 2020, doi: 10.1093/jamia/ocaa163.

[167] A. Gates, C. Johnson, and L. Hartling, "Technology-assisted title and abstract screening for systematic reviews: A retrospective evaluation of the Abstrackr machine learning tool," *Syst Rev*, vol. 7, no. 1, Mar. 2018, doi: 10.1186/s13643-018-0707-8.

[168] E. Guo, M. Gupta, J. Deng, Y.-J. Park, M. Paget, and C. Naugler, "Automated Paper Screening for Clinical Reviews Using Large Language Models," May 2023, doi: 10.2196/48996.

[169] A. Salinas and F. Morstatter, "The Butterfly Effect of Altering Prompts: How Small Changes and Jailbreaks Affect Large Language Model Performance," *ArXiv Computation and Language*, Jan. 2024, doi: 10.48550/arXiv.2401.03729.

[170] M. Michelson and K. Reuter, "The significant cost of systematic reviews and meta-analyses: A call for greater involvement of machine learning to assess the promise of clinical trials," *Contemp Clin Trials Commun*, vol. 16, Dec. 2019, doi: 10.1016/j.conctc.2019.100443.

[171] Z. Wang, T. Nayfeh, J. Tetzlaff, P. O'Blenis, and M. H. Murad, "Error rates of human reviewers during abstract screening in systematic reviews," Jan. 01, 2020, *Public Library of Science*. doi: 10.1371/journal.pone.0227742.

[172] J. A. Salvador-Oliván, G. Marco-Cuenca, and R. Arquero-Avilés, "Errors in search strategies used in systematic reviews and their effects on information retrieval," *Journal of the Medical Library Association*, vol. 107, no. 2, pp. 210–221, Apr. 2019, doi: 10.5195/jmla.2019.567.

[173] N. Buscemi, L. Hartling, B. Vandermeer, L. Tjosvold, and T. P. Klassen, “Single data extraction generated more errors than double data extraction in systematic reviews,” *J Clin Epidemiol*, vol. 59, no. 7, pp. 697–703, Jul. 2006, doi: 10.1016/j.jclinepi.2005.11.010.

[174] K. G. Shojania, M. Sampson, M. T. Ansari, J. Ji, S. Doucette, and D. Moher, “How Quickly Do Systematic Reviews Go Out of Date? A Survival Analysis,” 2007. [Online]. Available: www.annals.org

[175] G. Tsafnat, P. Glasziou, M. K. Choong, A. Dunn, F. Galgani, and E. Coiera, “Systematic review automation technologies,” Apr. 12, 2014, *BioMed Central Ltd.* doi: 10.1186/2046-4053-3-74.

[176] B. Tóth, L. Berek, L. Gulácsi, M. Péntek, and Z. Zrubka, “Automation of systematic reviews of biomedical literature: a scoping review of studies indexed in PubMed,” *Syst Rev*, vol. 13, no. 1, Dec. 2024, doi: 10.1186/s13643-024-02592-3.

[177] F. Chollet, *Deep Learning with Python*, 2nd ed. New York: Manning Publications, 2021.

[178] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient Estimation of Word Representations in Vector Space,” *ArXiv Computation and Language*, Jan. 2013, doi: 10.48550/arXiv.1301.3781.

[179] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks,” *ArXiv Computation and Language*, Aug. 2019, doi: 10.48550/arXiv.1908.10084.

[180] G. Salton and C. Buckley, “Term-weighting approaches in automatic text retrieval,” *Inf Process Manag*, vol. 24, no. 5, pp. 513–523, Jan. 1988, doi: 10.1016/0306-4573(88)90021-0.

[181] D. M. Blei, A. Y. Ng, and J. B. Edu, “Latent Dirichlet Allocation Michael I. Jordan,” 2003.

[182] A. K. Ambalavanan and M. V. Devarakonda, “Using the contextual language model BERT for multi-criteria classification of scientific articles,” *J Biomed Inform*, vol. 112, Dec. 2020, doi: 10.1016/j.jbi.2020.103578.

[183] B. K. Olorisade, P. Brereton, and P. Andras, “The use of bibliography enriched features for automatic citation screening,” *J Biomed Inform*, vol. 94, Jun. 2019, doi: 10.1016/j.jbi.2019.103202.

[184] K. E. K. Chai, R. L. J. Lines, D. F. Gucciardi, and L. Ng, “Research Screener: a machine learning tool to semi-automate abstract screening for systematic reviews,” *Syst Rev*, vol. 10, no. 1, Dec. 2021, doi: 10.1186/s13643-021-01635-3.

[185] T. Bekhuis and D. Demner-Fushman, “Towards automating the initial screening phase of a systematic review,” in *Studies in Health Technology and Informatics*, IOS Press, 2010, pp. 146–150. doi: 10.3233/978-1-60750-588-4-146.

[186] T. Bekhuis, E. Tseytlin, K. J. Mitchell, and D. Demner-Fushman, “Feature engineering and a proposed decision-support system for systematic reviewers of medical evidence,” *PLoS One*, vol. 9, no. 1, Jan. 2014, doi: 10.1371/journal.pone.0086277.

[187] Y. Mo, G. Kontonatsios, and S. Ananiadou, “Supporting systematic reviews using LDA-based document representations,” *Syst Rev*, vol. 4, no. 1, Nov. 2015, doi: 10.1186/s13643-015-0117-0.

[188] M. Miwa, J. Thomas, A. O’Mara-Eves, and S. Ananiadou, “Reducing systematic review workload through certainty-based screening,” *J Biomed Inform*, vol. 51, pp. 242–253, Oct. 2014, doi: 10.1016/j.jbi.2014.06.005.

[189] K. Hashimoto, G. Kontonatsios, M. Miwa, and S. Ananiadou, “Topic detection using paragraph vectors to support active learning in systematic reviews,” *J Biomed Inform*, vol. 62, pp. 59–65, Aug. 2016, doi: 10.1016/j.jbi.2016.06.001.

[190] J. C. Platt, “Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods,” *Advances in Large Margin Classifiers*, vol. 10, no. 3, pp. 61–74, 1999, [Online]. Available: https://www.researchgate.net/publication/2594015

[191] B. C. Wallace, K. Small, C. E. Brodley, J. Lau, and T. A. Trikalinos, “Deploying an Interactive Machine Learning System in an Evidence-Based Practice Center,” in *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, Jan. 2012. doi: 10.1145/2110363.211046.

[192] C. Bielza and P. Larrañaga, “Discrete Bayesian Network Classifiers,” *ACM Comput Surv*, vol. 47, no. 1, pp. 1–43, Jul. 2014, doi: 10.1145/2576868.

[193] K. C. Huang, I. J. Chiang, F. Xiao, C. C. Liao, C. C. H. Liu, and J. M. Wong, “PICO element detection in medical text without metadata: Are first sentences enough?,” *J Biomed Inform*, vol. 46, no. 5, pp. 940–946, Oct. 2013, doi: 10.1016/j.jbi.2013.07.009.

[194] S. Matwin, A. Kouznetsov, D. Inkpen, O. Frunza, and P. O’Blenis, “A new algorithm for reducing the workload of experts in performing systematic reviews,” *Journal of the American Medical Informatics Association*, vol. 17, no. 4, pp. 446–453, Jul. 2010, doi: 10.1136/jamia.2010.004325.

[195] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*. Wiley, 2013. doi: 10.1002/9781118548387.

[196] W. Shao *et al.*, “Aggregator: A machine learning approach to identifying MEDLINE articles that derive from the same underlying clinical trial,” *Methods*, vol. 74, pp. 65–70, Mar. 2015, doi: 10.1016/j.ymeth.2014.11.006.

[197] H. Almeida, M.-J. Meurs, L. Kosseim, and A. Tsang, “Data Sampling and Supervised Learning for HIV Literature Screening,” *IEEE Trans Nanobioscience*, vol. 15, no. 4, pp. 354–361, Jun. 2016, doi: 10.1109/TNB.2016.2565481.

[198] M. Ouzzani, H. Hammady, Z. Fedorowicz, and A. Elmagarmid, “Rayyan-a web and mobile app for systematic reviews,” *Syst Rev*, vol. 5, no. 1, Dec. 2016, doi: 10.1186/s13643-016-0384-4.

[199] P. Przybyła *et al.*, “Prioritising references for systematic reviews with RobotAnalyst: A user study,” *Res Synth Methods*, vol. 9, no. 3, pp. 470–488, Sep. 2018, doi: 10.1002/jrsm.1311.

[200] R. van de Schoot *et al.*, "An open source machine learning framework for efficient and transparent systematic reviews," *Nat Mach Intell*, vol. 3, no. 2, pp. 125–133, Feb. 2021, doi: 10.1038/s42256-020-00287-7.

[201] S. H. Cheng *et al.*, "Using machine learning to advance synthesis and use of conservation and environmental evidence," *Conservation Biology*, vol. 32, no. 4, pp. 762–764, Aug. 2018, doi: 10.1111/cobi.13117.

[202] Z. Yu, N. A. Kraft, and T. Menzies, "Finding better active learners for faster literature reviews," *Empir Softw Eng*, vol. 23, no. 6, pp. 3161–3186, Dec. 2018, doi: 10.1007/s10664-017-9587-0.

[203] I. J. Marshall, J. Kuiper, and B. C. Wallace, "RobotReviewer: Evaluation of a system for automatically assessing bias in clinical trials," *Journal of the American Medical Informatics Association*, vol. 23, no. 1, pp. 193–201, Jan. 2016, doi: 10.1093/jamia/ocv044.

[204] M. Issaiy *et al.*, "Methodological insights into ChatGPT's screening performance in systematic reviews," *BMC Med Res Methodol*, vol. 24, no. 1, Dec. 2024, doi: 10.1186/s12874-024-02203-8.

[205] X. Luo *et al.*, "Potential Roles of Large Language Models in the Production of Systematic Reviews and Meta-Analyses," *J Med Internet Res*, vol. 26, no. 1, 2024, doi: 10.2196/56780.

[206] R. Luo, Z. Sastimoglu, A. I. Faisal, and M. J. Deen, "Evaluating the Efficacy of Large Language Models for Systematic Review and Meta-Analysis Screening," *medrxiv*, 2024, doi: 10.1101/2024.06.03.24308405.

[207] W. E. Noreen, *Computer-Intensive Methods for Testing Hypotheses: An Introduction*. New York: Wiley, 1989.

[208] X. Bouthillier *et al.*, "Accounting for Variance in Machine Learning Benchmarks," *ArXiv Machine Learning*, Mar. 2021, doi: 10.48550/arXiv.2103.03098.

[209] S. Safari, A. Baratloo, M. Elfil, and A. Negida, "Evidence based emergency medicine; part 5 receiver operating curve and area under the curve," *Emergency*, vol. 4, no. 2, pp. 111–113, 2016, doi: 10.22037/aaem.v4i2.232.

[210] N. N. Pettit *et al.*, "Late onset infectious complications and safety of tocilizumab in the management of COVID-19," *J Med Virol*, vol. 93, no. 3, pp. 1459–1464, Mar. 2021, doi: 10.1002/jmv.26429.

[211] M. Gulersen *et al.*, "Histopathologic evaluation of placentas after diagnosis of maternal severe acute respiratory syndrome coronavirus 2 infection," *Am J Obstet Gynecol MFM*, vol. 2, no. 4, p. 100211, Nov. 2020, doi: 10.1016/j.ajogmf.2020.100211.

[212] L. Fadaizadeh, H. Jamaati, M. Varahram, M. J. Taheri, and M. Sanaat, "Follow-Up of Coronavirus Infected Patients Using Telemedicine in a Referral Pulmonary Center.," *Tanaffos*, vol. 19, no. 4, pp. 356–363, Dec. 2020.

[213] R. Vera-Pineda *et al.*, "Echocardiographic Characteristics of Subjects With COVID-19: A Case Series.," *Cardiol Res*, vol. 11, no. 4, pp. 260–265, Aug. 2020, doi: 10.14740/cr1084.

[214] H. Weerahandi *et al.*, "Post-Discharge Health Status and Symptoms in Patients with Severe COVID-19," *J Gen Intern Med*, vol. 36, no. 3, pp. 738–745, Mar. 2021, doi: 10.1007/s11606-020-06338-4.

[215] D. G. Campos *et al.*, "Screening Smarter, Not Harder: A Comparative Analysis of Machine Learning Screening Algorithms and Heuristic Stopping Criteria for Systematic Reviews in Educational Research," Mar. 01, 2024, *Springer*. doi: 10.1007/s10648-024-09862-5.

[216] A. Valizadeh *et al.*, "Abstract screening using the automated tool Rayyan: results of effectiveness in three diagnostic test accuracy systematic reviews," *BMC Med Res Methodol*, vol. 22, no. 1, Dec. 2022, doi: 10.1186/s12874-022-01631-8.

[217] A. Y. Tsou, J. R. Treadwell, E. Erinoff, and K. Schoelles, "Machine learning for screening prioritization in systematic reviews: Comparative performance of Abstrackr and EPPI-Reviewer," *Syst Rev*, vol. 9, no. 1, Apr. 2020, doi: 10.1186/s13643-020-01324-7.

[218] OpenAI, "GPT-4o Announcement," OpenAI. Accessed: Nov. 06, 2024. [Online]. Available: https://openai.com/index/hello-gpt-4o/

[219] OpenAI, "GPT-o1 Announcement," OpenAI. Accessed: Nov. 06, 2024. [Online]. Available: https://openai.com/index/introducing-openai-o1-preview/

[220] A. Paez, "Grey literature: An important resource in systematic reviews," *J Evid Based Med*, Dec. 2017, doi: 10.1111/jebm.12265.

[221] J. Schneider, L. Hoang, Y. Kansara, A. M. Cohen, and N. R. Smalheiser, "Evaluation of publication type tagging as a strategy to screen randomized controlled trial articles in preparing systematic reviews," *JAMIA Open*, vol. 5, no. 1, Apr. 2022, doi: 10.1093/jamiaopen/ooac015.

[222] Open Source, "React Native," Meta (Previously Facebook). Accessed: Nov. 30, 2024. [Online]. Available: https://github.com/facebook/react-native

[223] Open Source, "Expo," 650 Industries, Inc. Accessed: Nov. 30, 2024. [Online]. Available: https://github.com/expo/expo

[224] T. Neutkens, N. Kanezawa, G. Rauch, A. Susiripala, T. Kovanen, and D. Zajdband, "NextJS and Vercel," Vercel Inc. Accessed: Nov. 30, 2024. [Online]. Available: https://github.com/vercel/next.js/

[225] Open Source, "PrismaDB," Prisma Data, Inc. Accessed: Nov. 30, 2024. [Online]. Available: https://github.com/prisma/prisma

[226] MbientLab, "MetaMotionS Product Specification Datasheet," MbientLab. Accessed: Dec. 19, 2024. [Online]. Available: https://mbientlab.com/metamotions/

[227] M. Shanahan and B. Kalvin, “State of Mobile Internet Connectivity Report 2024,” GSMA: London, UK. Accessed: Jan. 09, 2025. [Online]. Available: https://www.gsma.com/r/somic/

[228] M. Straczkiewicz, P. James, and J. P. Onnela, “A systematic review of smartphone-based human activity recognition methods for health research,” *NPJ Digit Med*, vol. 4, no. 1, Dec. 2021, doi: 10.1038/s41746-021-00514-4.

[229] F. Demrozi, C. Turetta, F. Al Machot, G. Pravadelli, and P. H. Kindt, “A Comprehensive Review of Automated Data Annotation Techniques in Human Activity Recognition,” *ArXiv Machine Learning*, Jul. 2023, doi: 10.48550/arXiv.2307.05988.

[230] S. Hochreiter and J. Schmidhuber, “Long Short Term Memory,” *Neural Comput*, vol. 9, no. 8, Dec. 1997, doi: 10.1162/neco.1997.9.8.1735.

[231] K. Cho *et al.*, “Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation,” *ArXiv Computation and Language*, Jun. 2014, doi: 10.48550/arXiv.1406.1078.

[232] P. Bota, J. Silva, D. Folgado, and H. Gamboa, “A semi-automatic annotation approach for human activity recognition,” *Sensors (Switzerland)*, vol. 19, no. 3, Feb. 2019, doi: 10.3390/s19030501.

[233] H. Blunck, S. Bhattacharya, T. Prentow, M. Kjrgaard, and A. Dey, “Heterogeneity Activity Recognition Dataset,” UC Irvine Machine Learning Repository. Accessed: Nov. 30, 2024. [Online]. Available: https://archive.ics.uci.edu/dataset/344/heterogeneity+activity+recognition

[234] D. P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” *ArXiv Machine Learning*, Dec. 2014, doi: 10.48550/arXiv.1412.6980.

[235] N. Srivastava, G. Hinton, A. Krizhevsky, and R. Salakhutdinov, “Dropout: A Simple Way to Prevent Neural Networks from Overfitting,” *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, Jun. 2014, doi: 10.5555/2627435.2670313.

[236] S. Ioffe and C. Szegedy, “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift,” *ArXiv Machine Learning*, Feb. 2015, doi: 10.48550/arXiv.1502.03167.

[237] M. Ullrich *et al.*, “Detection of Unsupervised Standardized Gait Tests From Real-World Inertial Sensor Data in Parkinson’s Disease,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 29, pp. 2103–2111, 2021, doi: 10.1109/TNSRE.2021.3119390.

[238] J. Babayan *et al.*, “Everyday Life Tremor Signal Processing in PD Patients using BSN,” in *2021 Sixth International Conference on Advances in Biomedical Engineering (ICABME)*, IEEE, Oct. 2021, pp. 184–188. doi: 10.1109/ICABME53305.2021.9604898.

[239] J. R. Williamson, B. Telfer, R. Mullany, and K. E. Friedl, “Detecting Parkinson’s Disease from Wrist-Worn Accelerometry in the U.K. Biobank,” *Sensors*, vol. 21, no. 6, p. 2047, Mar. 2021, doi: 10.3390/s21062047.

[240] J. Varghese, A. Brenner, M. Fujarski, C. M. van Alen, L. Plagwitz, and T. Warnecke, "Machine Learning in the Parkinson's disease smartwatch (PADS) dataset," *NPJ Parkinsons Dis*, vol. 10, no. 1, p. 9, Jan. 2024, doi: 10.1038/s41531-023-00625-7.

[241] M. Nasseri *et al.*, "Signal quality and patient experience with wearable devices for epilepsy management," *Epilepsia*, vol. 61, no. S1, Nov. 2020, doi: 10.1111/epi.16527.

[242] D. Minici, G. Cola, G. Perfetti, S. Espinoza Tofalos, M. Di Bari, and M. Avvenuti, "Automated, ecologic assessment of frailty using a wrist-worn device," *Pervasive Mob Comput*, vol. 95, p. 101833, Oct. 2023, doi: 10.1016/j.pmcj.2023.101833.

[243] U. Großekathöfer *et al.*, "Automated Detection of Stereotypical Motor Movements in Autism Spectrum Disorder Using Recurrence Quantification Analysis," *Front Neuroinform*, vol. 11, Feb. 2017, doi: 10.3389/fninf.2017.00009.

[244] V. F. Rey, P. Hevesi, O. Kovalenko, and P. Lukowicz, "Let there be IMU data," in *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*, New York, NY, USA: ACM, Sep. 2019, pp. 699–708. doi: 10.1145/3341162.3345590.

[245] D. Balabka, "Semi-supervised learning for human activity recognition using adversarial autoencoders," in *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*, New York, NY, USA: ACM, Sep. 2019, pp. 685–688. doi: 10.1145/3341162.3344854.

[246] K. Takenaka, K. Kondo, and T. Hasegawa, "Segment-Based Unsupervised Learning Method in Sensor-Based Human Activity Recognition," *Sensors*, vol. 23, no. 20, p. 8449, Oct. 2023, doi: 10.3390/s23208449.

[247] X. Wu, Y. Zheng, C.-H. Chu, L. Cheng, and J. Kim, "Applying deep learning technology for automatic fall detection using mobile sensors," *Biomed Signal Process Control*, vol. 72, p. 103355, Feb. 2022, doi: 10.1016/j.bspc.2021.103355.

[248] E. El-Adawi, E. Essa, M. Handosa, and S. Elmougy, "Wireless body area sensor networks based human activity recognition using deep learning," *Sci Rep*, vol. 14, no. 1, p. 2702, Feb. 2024, doi: 10.1038/s41598-024-53069-1.

[249] E. Concha-Pérez, H. G. Gonzalez-Hernandez, and J. A. Reyes-Avendaño, "Physical Exertion Recognition Using Surface Electromyography and Inertial Measurements for Occupational Ergonomics," *Sensors*, vol. 23, no. 22, p. 9100, Nov. 2023, doi: 10.3390/s23229100.

[250] F. Zeng, M. Guo, L. Tan, F. Guo, and X. Liu, "Wearable Sensor-Based Residual Multifeature Fusion Shrinkage Networks for Human Activity Recognition," *Sensors*, vol. 24, no. 3, p. 758, Jan. 2024, doi: 10.3390/s24030758.

[251] O. Nafea, W. Abdul, G. Muhammad, and M. Alsulaiman, "Sensor-Based Human Activity Recognition with Spatio-Temporal Deep Learning," *Sensors*, vol. 21, no. 6, p. 2141, Mar. 2021, doi: 10.3390/s21062141.

[252] P. Pierleoni, L. Pernini, A. Belli, L. Palma, S. Valenti, and M. Paniccia, "SVM-based fall detection method for elderly people using Android low-cost smartphones," in *2015 IEEE Sensors Applications Symposium (SAS)*, IEEE, Apr. 2015, pp. 1–5. doi: 10.1109/SAS.2015.7133642.

[253] D. Gedminas, M. Dumpis, and A. Serackis, "Fusion of Activity Recognition and Recurrent Neural Network for Attitude Estimation Improvement," in *2022 IEEE Open Conference of Electrical, Electronic and Information Sciences (eStream)*, IEEE, Apr. 2022, pp. 1–4. doi: 10.1109/eStream56157.2022.9781733.

[254] D. Cheng, L. Zhang, L. Qin, S. Wang, H. Wu, and A. Song, "MaskCAE: Masked Convolutional AutoEncoder via Sensor Data Reconstruction for Self-Supervised Human Activity Recognition," *IEEE J Biomed Health Inform*, vol. 28, no. 5, pp. 2687–2698, May 2024, doi: 10.1109/JBHI.2024.3373019.

[255] X. Wang, Y. Lu, D. Wang, L. Liu, and H. Zhou, "Using Jaccard Distance Measure for Unsupervised Activity Recognition with Smartphone Accelerometers," 2017, pp. 74–83. doi: 10.1007/978-3-319-69781-9_8.

[256] Y. Wei, L. Liu, J. Zhong, Y. Lu, and L. Sun, "Unsupervised Race Walking Recognition Using Smartphone Accelerometers," 2015, pp. 691–702. doi: 10.1007/978-3-319-25159-2_63.

[257] Y.-L. Hsu *et al.*, "Design and Implementation of a Smart Home System Using Multisensor Data Fusion Technology," *Sensors*, vol. 17, no. 7, p. 1631, Jul. 2017, doi: 10.3390/s17071631.

[258] R. Delgado-Escano, F. M. Castro, J. R. Cozar, M. J. Marin-Jimenez, and N. Guil, "An End-to-End Multi-Task and Fusion CNN for Inertial-Based Gait Recognition," *IEEE Access*, vol. 7, pp. 1897–1908, 2019, doi: 10.1109/ACCESS.2018.2886899.

[259] S. U. Yunas and K. B. Ozanyan, "Gait Activity Classification Using Multi-Modality Sensor Fusion: A Deep Learning Approach," *IEEE Sens J*, vol. 21, no. 15, pp. 16870–16879, Aug. 2021, doi: 10.1109/JSEN.2021.3077698.

[260] J. Zhao and E. Obonyo, "Towards a Data-Driven Approach to Injury Prevention in Construction," 2018, pp. 385–411. doi: 10.1007/978-3-319-91635-4_20.

[261] A. Wang, G. Chen, C. Shang, M. Zhang, and L. Liu, "Human Activity Recognition in a Smart Home Environment with Stacked Denoising Autoencoders," 2016, pp. 29–40. doi: 10.1007/978-3-319-47121-1_3.

[262] N. Sarma, S. Chakraborty, and D. S. Banerjee, "Learning and Annotating Activities for Home Automation using LSTM," in *2019 11th International Conference on Communication Systems & Networks (COMSNETS)*, IEEE, Jan. 2019, pp. 631–636. doi: 10.1109/COMSNETS.2019.8711433.

[263] Y. Saez, A. Baldominos, and P. Isasi, "A Comparison Study of Classifier Algorithms for Cross-Person Physical Activity Recognition," *Sensors*, vol. 17, no. 1, p. 66, Dec. 2016, doi: 10.3390/s17010066.

[264] M. Nouredanesh, S. L. Kukreja, and J. Tung, "Detection of compensatory balance responses using wearable electromyography sensors for fall-risk assessment," in *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, IEEE, Aug. 2016, pp. 1680–1683. doi: 10.1109/EMBC.2016.7591038.

[265] Q. Cheng *et al.*, "Predicting Pulmonary Function from Phone Sensors.," *Telemed J E Health*, vol. 23, no. 11, pp. 913–919, Nov. 2017, doi: 10.1089/tmj.2017.0008.

[266] M. Abbas, D. Somme, and R. Le Bouquin Jeannès, "D-SORM: A digital solution for remote monitoring based on the attitude of wearable devices," *Comput Methods Programs Biomed*, vol. 208, p. 106247, Sep. 2021, doi: 10.1016/j.cmpb.2021.106247.

[267] O. Saidani, M. Alsafyani, R. Alroobaea, N. Alturki, R. Jahangir, and L. Jamel, "An Efficient Human Activity Recognition Using Hybrid Features and Transformer Model," *IEEE Access*, vol. 11, pp. 101373–101386, 2023, doi: 10.1109/ACCESS.2023.3314492.

[268] R. J. Adams, N. Saleheen, E. Thomaz, A. Parate, S. Kumar, and B. M. Marlin, "Hierarchical Span-Based Conditional Random Fields for Labeling and Segmenting Events in Wearable Sensor Data Streams.," *JMLR Workshop Conf Proc*, vol. 48, pp. 334–343, Jun. 2016.

[269] J. Andrés, G. Bailador, C. Gibbons, and A. Olmo, "Designing and Testing HealthTracker for Activity Recognition and Energy Expenditure Estimation within the DAPHNE Platform," *Procedia Comput Sci*, vol. 98, pp. 348–355, 2016, doi: 10.1016/j.procs.2016.09.052.

[270] D. Trabelsi, S. Mohammed, F. Chamroukhi, L. Oukhellou, and Y. Amirat, "An Unsupervised Approach for Automatic Activity Recognition based on Hidden Markov Model Regression," *ArXiv Machine Learning*, Dec. 2013, doi: 10.1109/TASE.2013.2256349.

[271] M. Munoz-Organero and R. Ruiz-Blazquez, "Time-Elastic Generative Model for Acceleration Time Series in Human Activity Recognition," *Sensors*, vol. 17, no. 2, p. 319, Feb. 2017, doi: 10.3390/s17020319.

[272] R. Fatima, M. H. Khan, M. A. Nisar, R. Doniec, M. S. Farid, and M. Grzegorzek, "A Systematic Evaluation of Feature Encoding Techniques for Gait Analysis Using Multimodal Sensory Data," *Sensors*, vol. 24, no. 1, p. 75, Dec. 2023, doi: 10.3390/s24010075.

[273] R. D. Riley, P. C. Lambert, and G. Abo-Zaid, "Meta-analysis of individual participant data: rationale, conduct, and reporting," *BMJ*, vol. 340, no. feb05 1, pp. c221–c221, Aug. 2010, doi: 10.1136/bmj.c221.

[274] I. J. Marshall and B. C. Wallace, "Toward systematic review automation: A practical guide to using machine learning tools in research synthesis," Jul. 11, 2019, *BioMed Central Ltd.* doi: 10.1186/s13643-019-1074-9.

[275] M. May, "Eight ways machine learning is assisting medicine," *Nat Med*, vol. 27, no. 1, pp. 2–3, Jan. 2021, doi: 10.1038/s41591-020-01197-2.

[276] A. H. Sapci and H. A. Sapci, "Artificial Intelligence Education and Tools for Medical and Health Informatics Students: Systematic Review," *JMIR Med Educ*, vol. 6, no. 1, p. e19285, Jun. 2020, doi: 10.2196/19285.

[277] M. Borenstein, J. P. T. Higgins, L. V. Hedges, and H. R. Rothstein, "Basics of meta-analysis: *I* 2 is not an absolute measure of heterogeneity," *Res Synth Methods*, vol. 8, no. 1, pp. 5–18, Mar. 2017, doi: 10.1002/jrsm.1230.

[278] N. Rieke *et al.*, "The future of digital health with federated learning," *NPJ Digit Med*, vol. 3, no. 1, p. 119, Sep. 2020, doi: 10.1038/s41746-020-00323-1.

[279] B. Haibe-Kains *et al.*, "Transparency and reproducibility in artificial intelligence," *Nature*, vol. 586, no. 7829, pp. E14–E16, Oct. 2020, doi: 10.1038/s41586-020-2766-y.

[280] Y. Wang *et al.*, "Guidelines, Consensus Statements, and Standards for the Use of Artificial Intelligence in Medicine: Systematic Review," *J Med Internet Res*, vol. 25, p. e46089, Nov. 2023, doi: 10.2196/46089.

[281] Y. Feng *et al.*, "Automated medical literature screening using artificial intelligence: a systematic review and meta-analysis," *Journal of the American Medical Informatics Association*, vol. 29, no. 8, pp. 1425–1432, Jul. 2022, doi: 10.1093/jamia/ocac066.

[282] R. Mc Ardle *et al.*, "Validation of an algorithm for detecting turning in people with cognitive impairment, considering dementia disease subtype," *Gait Posture*, vol. 118, pp. 141–147, May 2025, doi: 10.1016/j.gaitpost.2025.02.011.

[283] R. Mc Ardle *et al.*, "Gait in Mild Alzheimer's Disease: Feasibility of Multi-Center Measurement in the Clinic and Home with Body-Worn Sensors: A Pilot Study1," *Journal of Alzheimer's Disease*, vol. 63, no. 1, pp. 331–341, Apr. 2018, doi: 10.3233/JAD-171116.

[284] S. Šuster *et al.*, "Automating Quality Assessment of Medical Evidence in Systematic Reviews: Model Development and Validation Study," *J Med Internet Res*, vol. 25, p. e35568, Mar. 2023, doi: 10.2196/35568.

[285] J. Lees *et al.*, "Participant characteristics and exclusion from phase 3/4 industry funded trials of chronic medical conditions: meta-analysis of individual participant level data," *BMJ Medicine*, vol. 3, no. 1, p. e000732, May 2024, doi: 10.1136/bmjmed-2023-000732.

[286] A. Seth *et al.*, "OpenSim: Simulating musculoskeletal dynamics and neuromuscular control to study human and animal movement," *PLoS Comput Biol*, vol. 14, no. 7, p. e1006223, Jul. 2018, doi: 10.1371/journal.pcbi.1006223.

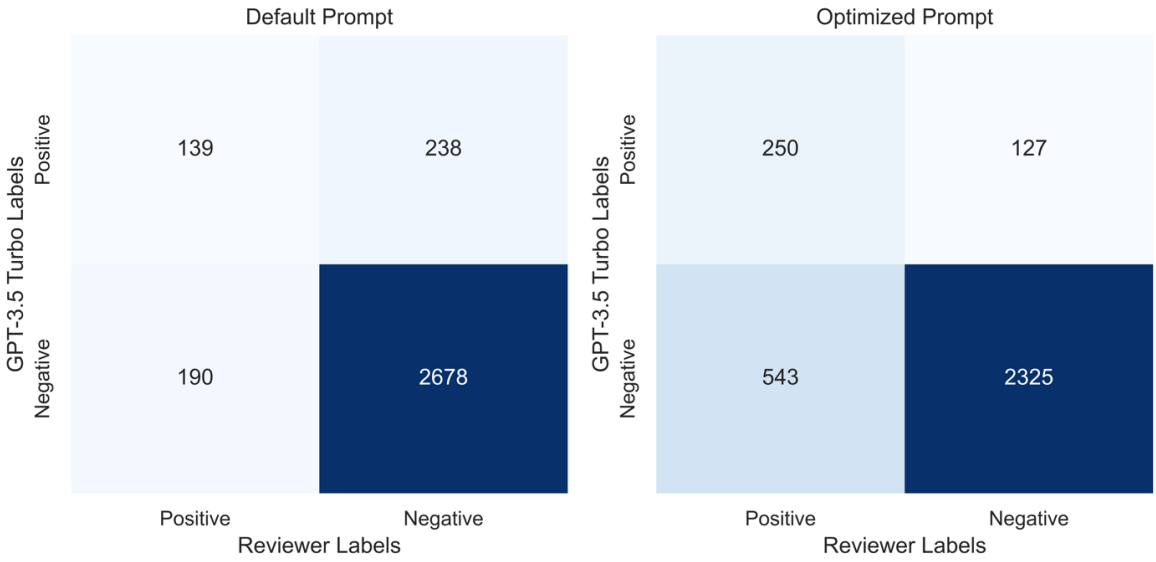
[287] X. Liu *et al.*, "Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension," *Lancet Digit Health*, vol. 2, no. 10, pp. e537–e548, Oct. 2020, doi: 10.1016/S2589-7500(20)30218-1.

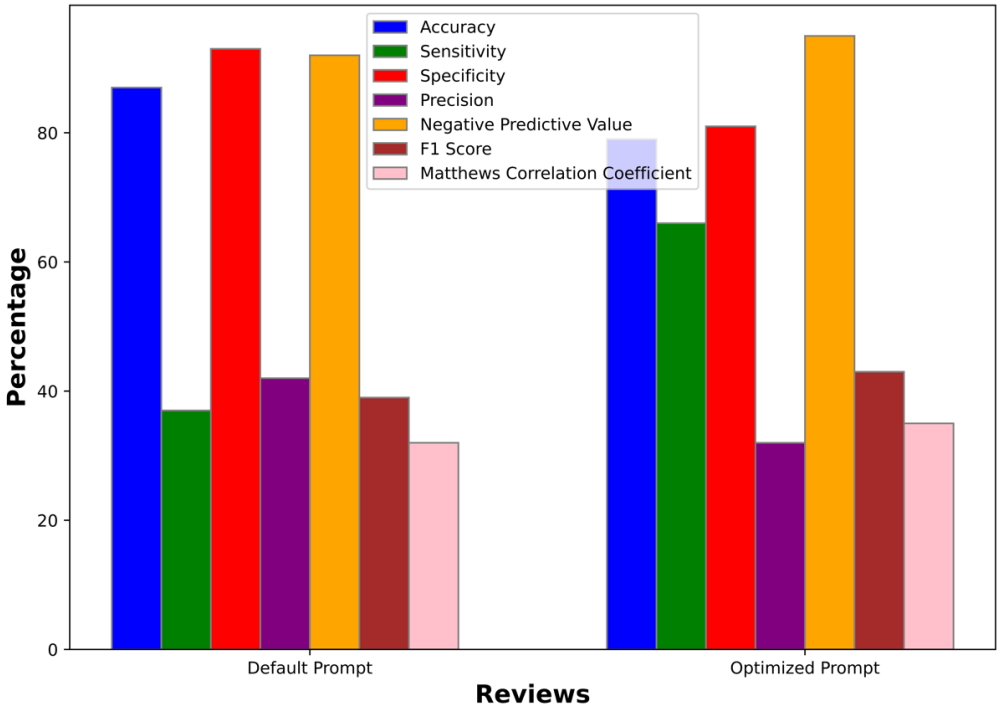
[288] A. B. Haque, A. K. M. N. Islam, and P. Mikalef, "Explainable Artificial Intelligence (XAI) from a user perspective: A synthesis of prior literature and problematizing avenues for future research," *Technol Forecast Soc Change*, vol. 186, p. 122120, Jan. 2023, doi: 10.1016/j.techfore.2022.122120.

[289] C. van Zyl, X. Ye, and R. Naidoo, "Harnessing eXplainable artificial intelligence for feature selection in time series energy forecasting: A comparative analysis of Grad-CAM and SHAP," *Appl Energy*, vol. 353, p. 122079, Jan. 2024, doi: 10.1016/j.apenergy.2023.122079.

[290] D. Trojaniello, A. Ravaschio, J. M. Hausdorff, and A. Cereatti, "Comparative assessment of different methods for the estimation of gait temporal parameters using a single inertial sensor: application to elderly, post-stroke, Parkinson's disease and Huntington's disease subjects," *Gait Posture*, vol. 42, no. 3, pp. 310–316, Sep. 2015, doi: 10.1016/j.gaitpost.2015.06.008.

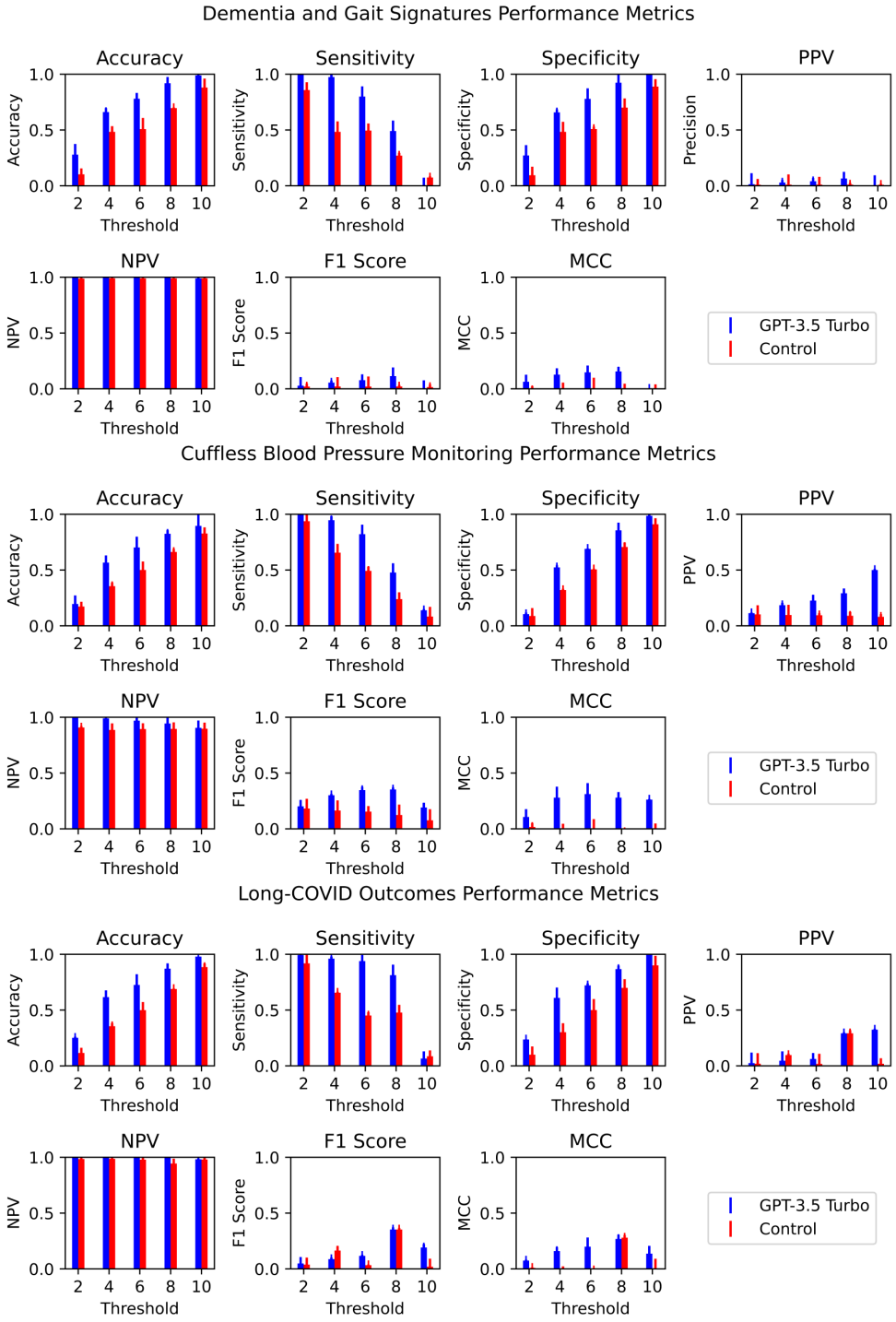
Appendix





Note: The optimized prompt, refined by removing overly stringent phrases from the original template, demonstrates a notable increase in true positives and a reduction in false negatives, leading to a doubled sensitivity in the title-abstract screening phase.

eFigure 1 Impact of Prompt Optimization on Model Performance: Enhanced Sensitivity in Dementia and Gait Signatures Review



Note: Bar charts show performance metrics like accuracy, sensitivity, and specificity for GPT-3.5 Turbo and a control group across different thresholds for three medical topics. Controls represent random multi-class classifier with bars denoting mean and standard deviation.

eFigure 2 Threshold Settings For GPT-3.5 Turbo Compared Against Random Multi-Class Control

Prompt Templates (Chapter 3 and 4)

1. Dementia and Gait Signatures

Evaluate the suitability of this study for inclusion in a meta-analysis focused on gait parameters in dementia. Studies should be observational, clinical, cross-sectional, or cohort, targeting Alzheimer's Disease (AD), Frontotemporal Dementia (FTD), Lewy Body Dementia (LBD), or Vascular Dementia (VaD). Exclude studies that are non-English publications, animal research, are review articles, books, conference abstracts, or lack a dementia population.

[1] or [2]

2. Cuffless Blood Pressure Monitoring

Evaluate the eligibility of this study for inclusion in a systematic review focused on cuffless BP monitoring technologies. Eligible studies must be based on Photoplethysmography (PPG), non-invasive, cuff-less, and tested against a reference device on human vital signs. Include devices and methods for BP estimation that are validated against standards such as ANSI/AAMI/ISO, with mean bias ≤ 5mm Hg and SD ≤ 8mm Hg for Systolic Blood Pressure (SBP) and Diastolic Blood Pressure (DBP) or have received an A or B grade from the European Hypertension Society (EHS). Additionally, studies should report a Mean Absolute Error (MAE) < 7mmHg, as per IEEE recommendations. Studies that achieved standard measurement accuracy in one database, supplemented by results from other databases, are also considered, even if those additional results fall outside standard ranges. Inclusion requires a minimum of 10 human participants.

[1] or [2]

3. Long-Covid Outcomes

Evaluate the eligibility of this study for inclusion in a systematic review focused on long COVID health outcomes. Eligible studies should report on health effects or symptoms that occur at least four weeks (28 days) after a COVID-19 diagnosis and must be peer-reviewed original research. Acceptable study designs include longitudinal, cohort, and cross-sectional studies that offer relevant data. Various methods for symptom measurement are permitted, including self-report, hospital equipment, and wearable technologies. Exclude meta-analyses, systematic reviews, case studies, and non-English publications due to translation barriers.

[1] or [2]

---

1 Should this study be included in the systematic review based on the criteria provided? Answer with True (for inclusion) or False (for exclusion)
2 On a scale of 1 to 10, where 1 is least likely and 10 is very likely, how suitable is this study for inclusion in the review based on the listed criteria?