Handgelenkbasierte Sensorsysteme zur Erkennung von Dyskinesien bei Parkinson - Chancen und Limitationen

Dissertation

zur Erlangung des akademischen Grades eines
Doktors der Medizin (Dr. med.)

an der

Medizinischen Fakultät der Universität Hamburg

vorgelegt von

Alexander Johannes Wiederhold

aus

Greven

2025

Betreuer:in / Gutachter:in der Dissertation: **Prof. Dr. med. Frank Ückert**

Gutachter:in der Dissertation: **Prof. Dr. med. Frank Ückert und Prof. Dr. med. Monika Pötter-Nerger**

Vorsitz der Prüfungskommission: **Prof. Dr. med. Monika Pötter-Nerger**

Mitglied der Prüfungskommission: **Prof. Dr. med. Renate Bonin-Schnabel**

Mitglied der Prüfungskommission: **Priv.-Doz. Dr. med. Christian Moll**

Datum der mündlichen Prüfung: **20.05.2026**

Inhaltsverzeichnis

1 Darstellung der Publikation

Diese Dissertation steht im Zusammenhang mit der Publikation „Opportunities and Limitations of Wrist-Worn Devices for Dyskinesia Detection in Parkinson's Disease", die 2025 im internationalen, peer-reviewten Fachjournal „Sensors" erschienen ist. Im Text genannte englischsprachige Abbildungen (Figures) und Tabellen (Tables) verweisen auf die zugrunde liegende englischsprachige Publikation. Entsprechende deutsche Bezüge finden sich in der Zusammenfassung.

1.1 Einleitung

Morbus Parkinson gehört zu den häufigsten neurodegenerativen Erkrankungen. Die Prävalenz liegt bei etwa 0,3 % der Bevölkerung über 40 Jahren und wird aufgrund der demografischen Alterung voraussichtlich weiter ansteigen (Pringsheim et al. 2014). Die klinische Versorgung von Parkinson-Patient:innen steht vor der Herausforderung, motorische Fluktuationen adäquat zu erfassen und zu therapieren. Besonders Levodopa-induzierte Dyskinesien (LID) erschweren mit fortschreitender Krankheitsdauer eine zielgerichtete Medikation (Kwon et al. 2022, Van Gerpen et al. 2006). Erforderlich ist eine engmaschige, individualisierte Titration, die starre Dosierungsschemata übersteigt und tageszeitliche sowie patientenspezifische Fluktuationen der Symptomatik berücksichtigt (Pringsheim et al. 2021). Die gängige klinische Praxis basiert meist auf punktuellen Beobachtungen im Rahmen ärztlicher Visiten, die tageszeitliche Schwankungen und interindividuelle Unterschiede nur unzureichend abbilden. Vor diesem Hintergrund rücken tragbare Sensorsysteme („Wearables") als Werkzeuge einer kontinuierlichen, objektiven Symptomerfassung zunehmend in den Fokus der Parkinson-Forschung (Giannakopoulou et al. 2022).

Während sich entsprechende Verfahren, beispielsweise die kontinuierliche Messung von Herzrhythmus oder Blutzuckerspiegel, in der Kardiologie und Diabetologie bereits etabliert sind (Bloem et al. 2023, Jalloul 2018), befindet sich die Implementierung einer kontinuierlichen Dyskinesiemessung in der Neurologie derzeit in einem frühen Entwicklungsstadium. Für die quantitative Erfassung von Dyskinesien existieren bislang nur begrenzt validierte Algorithmen, häufig auf Basis nicht-standardisierter Label oder heterogener Sensordaten. Studien wie jene

von Pfister et al. (2020) oder Hssayeni et al. (2021) berichten zwar von alltagsnahen Modellen mit Modellgenauigkeiten (Accuracy) im Bereich um 65 % bzw. r = 0,70–0,84, beruhen jedoch überwiegend auf nicht synchronisierten Beobachtungsratings. Dies bedeutet, dass die Messungen nicht durchgehend nach neurologischen Richtlinien gelabelt oder durch Expert:innen angeleitet wurden und eine erhebliche methodische Heterogenität aufweisen können. Die DREAM Challenge, als bislang größte öffentlich verfügbare Vergleichsstudie, erzielte bei der Dyskinesiedetektion lediglich eine Fläche unter der Precision Recall Kurve (AUPR) von 0,48, ein Indikator für die Herausforderungen bei der zuverlässigen Modellierung dieses Symptoms (Sieberts et al. 2021).

Weiterhin fehlt ein Konsens über die optimale Repräsentation von Beschleunigungsdaten und geeignete Evaluationsmetriken zur Bestimmung klinischer Anwendbarkeit. Umstritten ist insbesondere, ob Rohdaten in semantisch interpretierbare Bewegungsmerkmale überführt werden sollten oder ob eine vollständig datengetriebene automatisierte Zeitreihenanalyse vorzuziehen ist. Die beiden Ansätze unterscheiden sich hinsichtlich ihrer jeweiligen Stärken und Limitationen, insbesondere bezogen auf Interpretierbarkeit, Robustheit und die klinische Anschlussfähigkeit. Dabei erweist sich die Auswahl des Merkmalsraums ebenso wie die Festlegung der Evaluationsmetriken als entscheidend für die Leistungsfähigkeit der Modelle (Wang et al. 2022). Viele Arbeiten fokussieren sich auf die Modellgenauigkeit oder auf die AUPR, während klinisch relevante Kennzahlen wie Sensitivität und Spezifität häufig unzureichend berücksichtigt und nur selten systematisch miteinander verglichen werden (Hssayeni et al. 2021, Pfister et al. 2020).

Um diese Evidenzlücke zu schließen, braucht es kontinuierliche Sensorsignale, die unter klinisch kontrollierten Bedingungen erhoben und mit ärztlich verifizierten, standardisierten Labels verknüpft werden. Nur so lassen sich robuste, übertragbare Modelle für die Dyskinesiedetektion entwickeln.

Bevor jedoch maschinelle Lernverfahren für die Detektion von Dyskinesien eingesetzt werden können, sollte zunächst eine grundlegendere Herausforderung gelöst werden: Auf welche Weise können Dyskinesien algorithmisch überhaupt erfassbar gemacht werden und welche Aspekte des Bewegungssignals sind dabei aus klinischer Perspektive relevant? Dementsprechend setzt die Anwendung maschineller Lernmethoden voraus, ärztliche Beobachtungen systematisch in technische Repräsentationen zu überführen, die einerseits

klinisch interpretierbar und andererseits algorithmisch modellierbar sind. Von diesen Überlegungen ausgehend wurde die Annahme formuliert, dass eine semantische Datenrepräsentation, also eine an klinischen Kriterien orientierte Transformation von Zeitreihendaten, einen Mehrwert hinsichtlich Interpretierbarkeit und Generalisierbarkeit gegenüber rein datengetriebenen Ansätzen bieten könnte. Ergänzend dazu wurde hypothesiert, welche Leistungsfähigkeit ein vollständig datengetriebener Ansatz ohne domänenspezifisches Wissen auf Grundlage derselben Rohdaten erzielen könnte.

Die vorliegende Arbeit adressiert eben diese Fragestellungen: Sie validiert eine handgelenkbasierte Messmethodik im Routinebetrieb, prüft die Generalisierbarkeit von auf öffentlichen Datensätzen trainierten Modellen auf einen eigenständig erhobenen klinischen Datensatz und vergleicht zwei Repräsentationsstrategien: (1) einen semantisch fundierten Ansatz mit dimensionaler Reduktion und biomechanischen Merkmalen sowie (2) eine automatisierte, datengetriebene Zeitreihen-Merkmalsextraktion.

1.2 Material und Methoden

1.2.1 Daten

In dieser Arbeit wurden drei Bewegungsdatensätze verwendet: zwei öffentlich verfügbare Studien der Michael J. Fox Foundation sowie ein prospektiv erhobener klinischer Datensatz aus dem Universitätsklinikum Hamburg-Eppendorf (UKE).

1.2.2 Öffentliche Datensätze

Für die Entwicklung und initiale Modellvalidierung wurden zwei öffentlich verfügbare Bewegungsdatensätze der Michael J. Fox Stiftung (MJFF) verwendet: die Levodopa Response Study (LRS) und die Clinician Input Study (CIS-PD) (Motion Analysis Laboratory, D. O. P. M. und Robert 2019, Elm et al. 2019). Beide Studien zielen darauf ab, motorische Parkinson-Symptome und deren medikationsabhängige Fluktuationen mithilfe am Handgelenk getragener, triaxialer Beschleunigungssensoren zu erfassen. Die klinische Referenzbewertung erfolgte jeweils anhand der Unified Parkinson’s Disease Rating Scale (UPDRS).

Die LRS umfasst Messungen von 28 Parkinson-Patient:innen über vier konsekutive Tage hinweg. Am ersten und letzten Tag erfolgten strukturierte, klinische Untersuchungen inklusive standardisierter motorischer Testungen, wobei Patient:innen initial unter regulärer Medikation standen und am letzten Untersuchungstag nach vorherigem Absetzen der dopaminergen Medikation erneut untersucht wurden. Zwischen diesen klinischen Messungen wurden zwei Tage mit kontinuierlicher Datenerfassung im häuslichen Umfeld ergänzt (Motion Analysis Laboratory, D. O. P. M. und Robert 2019).

Die CIS-PD stellt eine sechsmonatige, longitudinale Studie mit insgesamt 51 Parkinson-Patient:innen dar, bei der periodische klinische Visiten mit standardisierten Untersuchungen erfolgten. In den Intervallen zwischen diesen Visiten trugen die Patient:innen kontinuierlich eine Smartwatch, ergänzten aber zusätzlich subjektive Symptom- und Medikationsberichte über eine App (Elm et al. 2019).

In der vorliegenden Arbeit werden beide Datensätze zusammengeführt und als ein kombinierter „MJFF Datensatz“ referenziert.

1.2.3 Eigene Datenerhebung (PACMAN-Datensatz)

Die Erhebung des PACMAN-Datensatzes (Parkinson Clinical Movement Assessment) wurde im Rahmen einer nicht-interventionellen klinischen Studie über vier Monate eigenständig geplant und durchgeführt. Die Datensammlung fand im strukturierten Umfeld der Parkinson Komplexbehandlung (PKB) an der Klinik für Neurologie des UKE statt. Die PKB ist laut Richter et al. ein mehrtägiges, interdisziplinäres stationäres Behandlungsprogramm mit standardisierter neurologischer Diagnostik und regelmäßigen therapeutischen Einheiten, welches eine tägliche ärztliche Beobachtung ermöglicht (Richter et al. 2019). Durch die Einbettung der Datenerhebung in diesen klinischen Rahmen entstand ein kontrolliertes Setting, in dem kontinuierliche sensorische Rohdaten eng mit ärztlich validierten motorischen Bewertungen verknüpft werden konnten.

An jedem Behandlungstag wurden mindestens zwei klinische Visiten durchgeführt. Zu Beginn des Aufenthalts erhielten die Patient:innen eine Smartwatch (Apple Watch Series 6), deren korrekte Positionierung, Achsenausrichtung und Passform täglich durch den Doktoranden überprüft und dokumentiert wurde. Die Smartwatch verblieb anschließend durchgehend am stärker betroffenen Handgelenk und zeichnete triaxiale Beschleunigungsdaten kontinuierlich

mit hoher zeitlicher Auflösung auf. Bei jeder klinischen Visite erfolgte eine standardisierte, motorische Untersuchung gemäß des dritten Teils der UPDRS (Teil III). Für die Bearbeitung der in dieser Arbeit formulierten Forschungsfrage wurde das UPDRS-Item 3.6 („alternierende Handbewegungen“) ausgewählt, da sich diese regelmäßig periodische Bewegungsaufgabe besonders für eine semantische Datenrepräsentation eignet und eine valide Grundlage zur Extraktion klinisch relevanter Merkmale bietet. Die entsprechenden Untersuchungen wurden unmittelbar nach den UPDRS-Kriterien bewertet und mit präzisen Zeitstempeln annotiert. Diese zeitlich definierten Bewertungen dienten anschließend als Referenzlabels für die algorithmische Datenverarbeitung.

Die gesamte Datenerhebung war vollständig in die klinische Routine integriert und erfolgte gemäß der Deklaration von Helsinki. Alle Studienteilnehmer:innen gaben vor Einschluss eine schriftliche Einwilligung ab. Das Studienprotokoll wurde vorab durch die Ethikkommission der Ärztekammer Hamburg positiv bewertet (ID: 2022-100846-BO-ff).

1.2.4 Einheitliche Dateninfrastruktur und methodische Vorverarbeitung

Für diese Arbeit wurde eine zuvor entwickelte, umfassende Datenarchitektur genutzt, die im Rahmen eines vorausgehenden Projekts des UKE konzipiert und implementiert wurde (Gundler et al. 2023). Diese Architektur basiert auf dem international etablierten HL7 FHIR-Standard (Fast Healthcare Interoperability Resources) und ermöglicht den interoperablen Austausch sowie die standardisierte Speicherung klinischer Bewegungsdaten.

Als methodische Weiterentwicklung wurde das Datenhandling der unterschiedlich langen Messsequenzen angepasst. Übliche Zeitreihenanalysen verwenden häufig eine Segmentierung („Windowing“) der Sensordaten in feste Fensterlängen, riskieren jedoch bei klinisch annotierten Bewegungsaufgaben Informationsverluste durch Abschneiden oder Verzerrungen relevanter Signalabschnitte (Banos et al. 2014, Shawen et al. 2020). Daher wurde hier bewusst auf feste Fenstergrößen verzichtet und stattdessen jede Aufzeichnung in ihrer vollständigen Dauer gespeichert und verarbeitet, damit die semantische Einheit zwischen Bewegungssignal und ärztlicher Annotation auch als solche repräsentiert wird.

Zur Sicherstellung der Vergleichbarkeit zwischen unterschiedlichen Geräten, individuellen Trageweisen und Körperseiten wurde außerdem eine globale Rotationsnormalisierung angewandt. Diese Normalisierung kompensiert systematische

Abweichungen der räumlichen Ausrichtung der Sensorachsen und verhindert somit lageabhängige Verzerrungen in den nachfolgenden Analysen.

1.2.5 Datenrepräsentationen als Brücke zwischen klinischer Beobachtung und algorithmischer Analyse

Beschleunigungssignale sind in ihrer Rohform numerisch und visuell kaum intuitiv verständlich. Klinische Beurteilungen orientieren sich hingegen an Qualitäten wie Rhythmus, Amplitude und Symmetrie. Methodisch besteht daher die Aufgabe, diese wahrnehmbaren Eigenschaften als technische Merkmale abzubilden und die klinische Semantik zu erhalten. Unter Datenrepräsentation versteht man die gezielte Transformation hochdimensionaler Sensorzeitreihen in strukturierte Merkmalsräume, in denen informationshaltige Muster extrahierbar bleiben, ohne wesentliche Verluste zu erzeugen (Bengio et al. 2014). So werden die Daten zugleich für Fachpersonal interpretierbar und für algorithmische Verfahren verarbeitbar.

1.2.6 Semantische Repräsentation

Die semantische Repräsentation verfolgt das Ziel, klinische Beobachtungen von motorischen Parkinson-Symptomen systematisch in technisch interpretierbare Merkmale zu übersetzen.

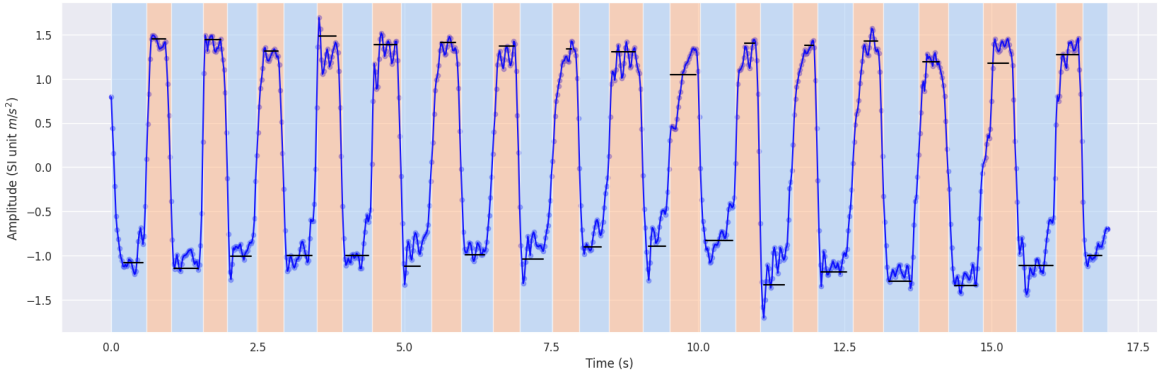

Abbildung 1: *Beispielhaft segmentierte Messung. Die blau markierten Segmente besitzen ein Extremum < 0, wohingegen die orange markierten Segmente ein Extremum > 0 aufweisen. Diese Segmentierung wird für die anschließende Merkmalsanalyse genutzt.*

Ausgangspunkt ist das in Abschnitt 1.2.3 definierte Bewegungsitem (alternierende Handbewegungen), welches aufgrund seines reproduzierbaren, rhythmischen Charakters

besonders geeignet ist, um algorithmisch verwertbare Signalcharakteristika zu extrahieren. Es wird folglich angenommen, dass diese standardisierte motorische Aufgabe, im Gegensatz zu alltäglichen, variablen Bewegungen, stabile Schwingungsmuster aufweist, die sowohl visuell nachvollziehbar als auch rechnerisch klar abgrenzbar sein könnten.

Zur methodischen Komplexitätsreduktion wurde zunächst eine eindimensionale Projektion des ursprünglichen tri-axialen Beschleunigungssignals durchgeführt. Dabei kam eine Hauptkomponentenanalyse (PCA) zum Einsatz, welche die ursprünglichen Sensordaten entlang ihrer Varianzachse reduziert. Das Ergebnis dieser Transformation bildet eine übersichtliche, eindimensionale Repräsentation des dominanten Bewegungsmusters, welche die relevanten Bewegungsinformationen bewahrt und zugleich eine einfache visuelle Plausibilitätsprüfung (etwa hinsichtlich rhythmischer Stabilität) ermöglicht. Das Transformationsergebnis ist für eine exemplarische Messung in Abbildung 1 dargestellt.

Die anschließende Merkmalsextraktion orientiert sich methodisch an der Arbeit von Sánchez-Fernández et al. (2023), welche eine umfassende Palette biomechanischer Merkmale für die Charakterisierung alternierender Handbewegungen vorstellten. Für die vorliegende Arbeit erfolgte jedoch eine eigenständige, gezielte Anpassung und Auswahl dieser Merkmale, da hier ausschließlich Beschleunigungsdaten zur Verfügung standen. Von ursprünglich 20 vorgeschlagenen Merkmalen wurden so insgesamt neun Merkmale ausgewählt. Zu diesen Merkmalen zählen unter anderem absolute Mittelwerte, relative Maxima und Minima der Mittelwerte einzelner Abschnitte sowie Dauerverhältnisse dieser Intervalle. Eine ausführliche Darstellung dieser Merkmale findet sich in der zugehörigen Originalpublikation.

Zur detaillierten dynamischen Analyse wurde jede Bewegungsaufzeichnung in drei gleich große zeitliche Abschnitte (Anfang, Mitte, Ende) unterteilt. Diese Segmentierung ermöglicht es, Merkmale nicht nur auf Gesamtsequenzen, sondern auch spezifisch auf kleinere, inhaltlich zusammenhängende Teilabschnitte anzuwenden. Dadurch können zeitlich lokale Veränderungen innerhalb der Bewegung präziser identifiziert und bewertet werden. Visuell interpretierbare Projektionen und domänenspezifische Merkmalsextraktion transformieren so die ärztliche Beobachtungslogik in algorithmisch verwertbare Strukturen.

1.2.7 Automatisierte Merkmalsextraktion

Im Gegensatz zur semantischen Repräsentation, die klinisches Wissen direkt integriert, verfolgt die automatisierte Merkmalsextraktion einen rein datengetriebenen, algorithmischen Ansatz. Ziel dieser Methode ist es, ohne domänenspezifische Vorauswahl möglichst viele potenziell relevante Merkmale automatisiert zu generieren, um auch klinisch nicht unmittelbar erkennbare Muster in den Sensordaten zu erfassen.

Hierzu kam in dieser Arbeit die Python-Bibliothek tsfresh (Version 0.20.1) zum Einsatz, welche algorithmisch und vollständig automatisiert mehr als 3000 Merkmale aus den Bereichen Statistik, Zeitreihenanalyse und Signalverarbeitung extrahiert (Christ et al. 2018). Als Eingangsdaten wurden entweder die originalen tri-axialen Beschleunigungssignale oder deren eindimensionale PCA-Projektionen verwendet. Das Ergebnis dieses automatisierten Prozesses ist ein hochdimensionaler Merkmalsraum aus statistischen Kennwerten, Zeitreihencharakteristika und signalanalytischen Koeffizienten, der eine Grundlage für die spätere Modellierung motorischer Dynamiken bietet. Entsprechend ergeben sich drei verschiedene Repräsentationsstrategien:

1. Semantische Pipeline (PCA + manuell ausgewählte biomechanische Merkmale),
2. PCA-Projektion kombiniert mit automatisierter Merkmalsextraktion,
3. Automatisierte Merkmalsextraktion basierend auf originalen, dreidimensionalen Rohsignalen.

1.2.8 Modelltraining und Evaluation

Für die Klassifikation der Dyskinesie-Schweregrade (UPDRS-Item 3.6) wurden aufbauend auf den erzeugten Merkmalsrepräsentationen Verfahren des maschinellen Lernens eingesetzt. Dabei wurde eine systematische Hyperparameteroptimierung (Liashchynskyi und Liashchynskyi 2019) auf Basis des zusammengeführten MJFF-Datensatzes durchgeführt. Dieser Prozess setzte sich aus einer Merkmalsselektion, eines Klassengrößenausgleichs und eines Klassifikator-Vergleichs mit Hyperparameteroptimierung zusammen. Details zu den verwendeten Modellen und deren Konfigurationen finden sich in der Originalarbeit.

Zur Evaluation wurde der ungewichtete F_1-Score verwendet, definiert als das harmonische Mittel aus Präzision und Sensitivität. Diese Metrik berücksichtigt sowohl falsch-positive als auch falsch-negative Klassifikationen und eignet sich besonders für Szenarien mit

unausgeglichener Klassenverteilung (Seo et al. 2021). In Ermangelung eines klinisch standardisierten Einzelwertes zur Beurteilung diagnostischer Güte, bietet der F_1-Score eine ausgewogene Bewertung der Klassifikationsleistung. (Hssayeni et al. 2021, Pfister et al. 2020). Zur robusten Validierung der Modellleistungen erfolgte anschließend eine unabhängige Evaluation auf den eigenständig erhobenen PACMAN-Datensatz. Die jeweils zehn besten Modelle jeder Repräsentationsstrategie wurden ohne weitere Anpassungen auf den PACMAN-Datensatz angewendet und evaluiert. Die methodische Vergleichbarkeit wurde dabei sichergestellt, indem dieselben Bewertungsmethoden und statistischen Testverfahren (Welch-korrigierter t-Test, Signifikanzniveau $p < 0{,}001$) wie in der initialen Entwicklung genutzt wurden.

1.3 Ergebnisse

1.3.1 Eingeschlossene Patient:innenkohorten und Datenintegrität

Die LRS-Kohorte umfasst 27 Parkinson-Patient:innen (Alter: 67 ± 9 Jahre), mit durchschnittlich 51 ± 9 validen, ärztlich annotierten Messungen. Die CIS-PD-Kohorte besteht aus 24 Patient:innen (Alter: 63 ± 10 Jahre), mit durchschnittlich 12 ± 3 Messungen pro Person. Die eigenständig erhobene PACMAN-Kohorte umfasst 25 Patient:innen (Alter: 65 ± 8 Jahre), wobei im Durchschnitt 3 ± 2 Messungen pro Patient:in unter engmaschiger klinischer Aufsicht erhoben wurden.

Die Datenintegrität wurde methodisch durch gezielte Auswahl ausschließlich klinisch beaufsichtigter und standardisiert annotierter Messungen sichergestellt. Unbeaufsichtigte, häusliche Messungen wurden nicht berücksichtigt, um eine hohe Validität und methodische Konsistenz zwischen den Datensätzen sicherzustellen. Die verwendete Sensorik sowie standardisierte neurologische Bewertung anhand der international anerkannten UPDRS-Richtlinien gewährleisteten zudem eine hohe methodische Vergleichbarkeit zwischen den Kohorten.

1.3.2 Modellleistung auf den Trainingsdaten

Zur Bewertung der Vorhersagegüte der entwickelten Klassifikationsmodelle auf den Trainingsdaten wurde das MJFF-Datenset herangezogen. Abbildung 2 zeigt die

Konfusionsmatrix der jeweils leistungsstärksten Modellkonfiguration für jede der drei gewählten Repräsentationsmethoden. Diese Matrix erlaubt eine differenzierte Gegenüberstellung der tatsächlichen Label (klinisch vergebene Dyskinesie-Einstufung) mit den vom Modell vorhergesagten Klassen. Alle drei Modelle erkannten die Abwesenheit von Dyskinesien zuverlässig.

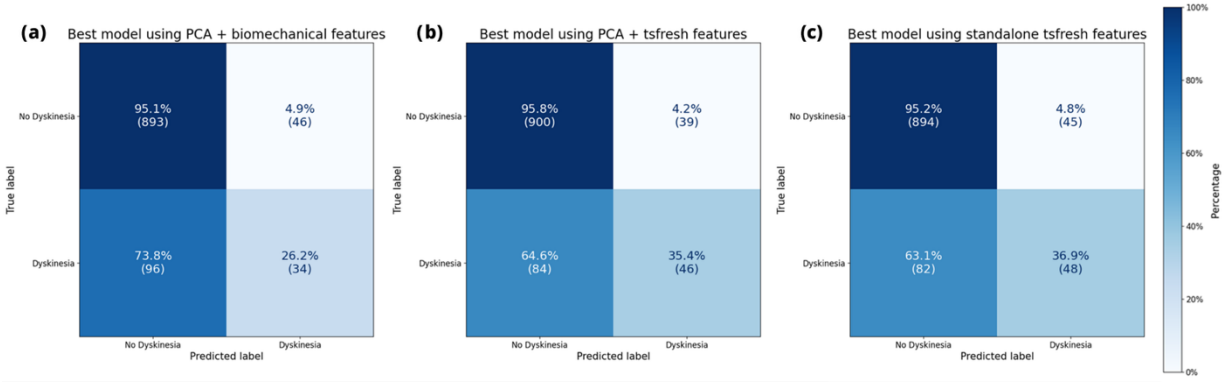

Abbildung 2: *Konfusionsmatrix der leistungsstärksten Modelle für jedes Repräsentationsverfahren auf dem MJFF-Datensatz.*

Bei der Identifikation tatsächlicher Dyskinesie-Episoden erzielten alle drei Repräsentationsverfahren vergleichbare Ergebnisse. Die automatisierte Merkmalsextraktion (mit und ohne vorherige PCA) erreichte dabei die höchsten maximalen F_1-Werte von jeweils 0,68. Die semantische Repräsentation erzielte mit einem maximalen F_1-Score von 0,63 eine leicht niedrigere Leistung. Tabelle 1 zeigt die jeweils beste Modellkonfiguration für jedes Repräsentationsverfahren auf dem Trainingsdatensatz. Statistische Analysen (siehe 1.2.8) zeigten keine signifikanten Unterschiede zwischen den drei methodischen Ansätzen im Trainingssetting.

Tabelle 1. Ergebnisse der leistungsstärksten Modelle für jede Repräsentationskonfiguration auf dem MJFF-Datensatz

Repräsentation	**Modell**			**Macro F_1**	**Genauig-**
	Klassifikator	**Selektor**	**Sampling**	**Wert**	**keit**
PCA + biomechanische Merkmale	gradient boosted trees (lr = 1.0)	10	ohne	0.63	0.87
PCA + tsfresh kombiniert	random forest	ohne	SMOTE	0.68	0.88
tsfresh alleinstehend	random forest	10	ohne	0.68	0.88

lr = Lernrate, SMOTE = Synthetic Minority Over-sampling Technique.

1.3.3 Modellleistung auf den PACMAN-Daten

Zur externen Validierung und methodischen Prüfung der klinischen Generalisierbarkeit wurden die jeweils zehn leistungsstärksten Modelle aus dem MJFF-Datensatz unverändert auf den eigenständig erhobenen PACMAN-Datensatz übertragen. Abbildung 3 zeigt die Konfusionsmatrix des jeweils besten Modells pro Methode. Die Klassifikationsmuster zeigen, dass die Erkennung der Dyskinesie-freien Zustände zuverlässig gelingt, während die Identifikation tatsächlicher Dyskinesien mit geringerer Genauigkeit erfolgt.

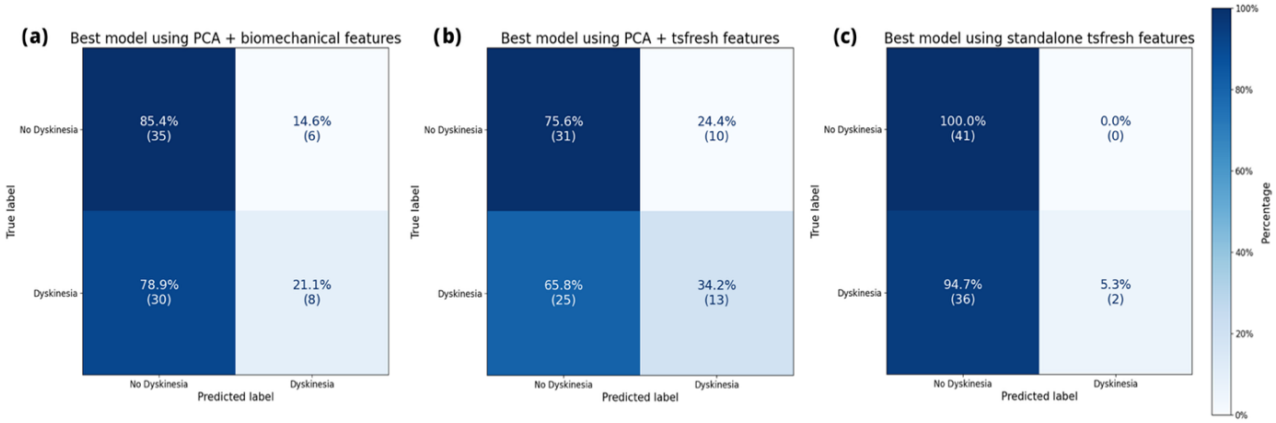

Abbildung 3: *Konfusionsmatrix der leistungsstärksten Modelle für jedes Repräsentationsverfahren auf dem PACMAN-Datensatz.*

In diesem klinischen Validierungsschritt zeigte sich eine signifikante Leistungsdifferenz zwischen den verschiedenen Repräsentationsansätzen. Die höchste Generalisierbarkeit (F_1-Score = 0,53) wurde durch die Kombination aus PCA und automatisierter Merkmalsextraktion erreicht. Die semantische Methode folgte knapp dahinter ($F_1 = 0{,}48$). Deutlich geringer schnitt das rein automatisierte Verfahren auf Originaldatenbasis ab ($F_1 = 0{,}37$). Die statistische Analyse bestätigte signifikante Unterschiede: Die semantische Methode war signifikant besser als das vollständig automatisierte Verfahren ($p < 0{,}001$).

Tabelle 2. Ergebnisse der leistungsstärksten Modelle für jede Repräsentationskonfiguration auf dem PACMAN-Datensatz

Repräsentation	Modell			Macro F_1	Genauig-
	Klassifikator	Selektor	Sampling	Wert	keit
PCA + biomechanische Merkmale	gradient boosted trees (lr = 1.0)	ohne	SMOTE	0.48	0.54
PCA + tsfresh kombiniert	gradient boosted trees (lr = 1.0)	10	ohne	0.53	0.56
tsfresh alleinstehend	k-nearest neighbors (nn = 5)	10	ohne	0.37	0.54

lr = Lernrate, nn = Anzahl benachbarter Knoten, SMOTE = Synthetic Minority Over-sampling Technique.

1.3.4 Einordnung in den klinischen Kontext

Im Ergebnis zeigte sich, dass eine dimensionsreduzierte Projektion der Rohdaten, unabhängig von der spezifischen Merkmalsextraktion, entscheidend zu den bestplatzierten Modellen beitrug. Während die automatisch extrahierten Merkmale im internen Trainingssetting leicht bessere Ergebnisse lieferten (maximaler F_1-Score 0,68 vs. semantisch 0,63), zeigte sich in der externen Validierung auf PACMAN ein differenziertes Bild: Die hybride Kombination aus PCA und automatischer Merkmalsextraktion schnitt mit einem F_1-Score von 0,53 am besten ab, gefolgt von der semantischen Methode (F_1 = 0,48), während die rein automatische Merkmalsextraktion deutlich abfiel (F_1 = 0,37).

1.4 Diskussion

Die vorangehende Einordnung zeigt, dass klinisch begründete Signalverarbeitungsschritte und dimensionsreduzierte Projektionen die Generalisierbarkeit unter realen klinischen Bedingungen erhöhen und rein datengetriebenen Merkmalsextraktionen insbesondere in heterogenen Settings überlegen sind. Diese Einschätzung wird durch die externe Validierung gestützt: Die Diskrepanz zwischen internem Training (MJFF) und der Leistung auf PACMAN markiert die Grenzen rein datengetriebener Ansätze. Zugleich unterstreicht sie die zentrale Bedeutung standardisierter klinischer Bedingungen und ärztlich validierter Label als Voraussetzung für

zuverlässige und übertragbare Modelle. Öffentliche Datensätze bieten zwar methodisch wertvolle Ansätze zur initialen Modellentwicklung, doch erst klinisch kontrollierte Erhebungen wie PACMAN ermöglichen belastbare Aussagen zur Anwendbarkeit in realen klinischen Settings.

Die Wahl geeigneter Evaluationsmetriken ist entscheidend für eine sinnvolle und vergleichbare Einordnung der Ergebnisse. Zwei der bislang größten Studien zur Detektion von Dyskinesien veranschaulichen die damit verbundenen Herausforderungen: Sowohl Hssayeni et al. (2021) als auch Pfister et al. (2020) bewerten ihre Modelle vorrangig anhand der Modellgenauigkeit sowie des Korrelationskoeffizienten *r* (Hssayeni et al. 2021, Pfister et al. 2020). Obwohl die Modellgenauigkeit bei ausgewogen verteilten Klassen als geeignete Metrik gilt, kann sie in Fällen starker Klassenungleichgewichte zu überoptimistischen Einschätzungen der Modellgüte führen. Klinisch erhobene Datensätze zeichnen sich jedoch typischerweise durch begrenzte Stichprobengrößen und eine ungleiche Verteilung der Annotationen aus, da der Fokus auf qualitativ hochwertigen und standardisierten Messungen liegt (Seo et al. 2021). Der ungewichtete F_1-Score stellt in solchen Kontexten eine robustere Alternative dar, da er die klinische Relevanz falsch-positiver und falsch-negativer Vorhersagen explizit berücksichtigt. So wird erneut die Notwendigkeit standardisierter Bewertungsmetriken in der Dyskinesiedetektion unterstrichen. Im Gegensatz zur Modellgenauigkeit, welche insbesondere bei Klassenungleichgewicht nur bedingt aussagekräftig ist, bietet der F_1-Score einen klinisch differenzierenden Maßstab. Die beobachteten Ergebnisse (F_1 = 0,68 Trainingsset, F_1 = 0,53 externe Validierung) liegen im Rahmen vergleichbarer Literaturwerte, jedoch erschweren Unterschiede in Studiendesign und Metriken einen unmittelbaren Vergleich (Hssayeni et al. 2021, Pfister et al. 2020). Dies wird insbesondere durch die in Abbildung 3 und 4 dargestellten Konfusionsmatrizen verdeutlicht: Sowohl im Trainings- als auch im Validierungsdatensatz werden dyskinesiefreie Zustände mit höherer Zuverlässigkeit erkannt, während die Erkennung tatsächlicher Dyskinesien eine geringere Genauigkeit zeigt.

Da der F_1-Score die Balance von Präzision und Sensitivität unter Klassenungleichgewicht abbildet, verweist das beobachtete Leistungsniveau weniger auf reine Modellkapazität als auf die Güte der zugrunde liegenden Datenrepräsentation. Somit wird das Fehlerprofil wesentlich durch die von der Repräsentationsstrategie exponierten Signalstrukturen bestimmt. Die Wahl der geeigneten Repräsentationsstrategie ist folglich

entscheidend, um die Generalisierbarkeit der Modelle über heterogene klinische Settings hinweg beurteilen zu können.

Die datengetriebene Natur der automatisierten Extraktion erlaubt zwar die Erkennung bislang unbekannter Muster, geht jedoch auf Kosten klinischer Interpretierbarkeit und erhöht das Risiko eines Übertrainierens (Overfitting), insbesondere bei kleinen oder heterogenen Stichproben. Demgegenüber überführt die semantische Methode klinische Beobachtung explizit in algorithmisch verwertbare Parameter, wodurch sie klinisch transparenter und gegenüber methodischen Artefakten robuster ist.

Die Hybridstrategie kombiniert die jeweilige methodische Stärke beider Ansätze: klinisch-motivierte Dimensionalitätsreduktion zur kontrollierten Datenkomprimierung und automatische Extraktion zur Erfassung residualer Muster. Die Ergebnisse unterstreichen erstmals systematisch, dass ein solcher hybrider Ansatz am ehesten geeignet erscheint, sowohl klinischen Anforderungen an Transparenz und Interpretierbarkeit als auch algorithmischen Anforderungen an robuste Mustererkennung gerecht zu werden.

Im Hinblick auf die klinische Implementierung lassen sich aus den Befunden drei Anforderungen ableiten: (1) standardisierter motorischer Tasks mit klarer klinischer Verankerung, (2) transparenter, reproduzierbarer Datenrepräsentationen und (3) metrischer Harmonisierung für Vergleichbarkeit. Aus klinischer Sicht eröffnet die hier beschriebene methodische Kombination eine vielversprechende Perspektive für den Einsatz in Entscheidungsunterstützungssystemen zur Anpassung dopaminerger Therapien. Insbesondere könnte die Integration in klinische Versorgungspfade (z. B. automatisiertes Monitoring im Rahmen stationärer Therapien) die frühzeitige und präzisere Anpassung medikamentöser Dosierungen erleichtern und damit langfristig zu verbesserten Patientenverläufen beitragen.

Die vorliegenden Ergebnisse unterliegen mehreren Einschränkungen. Erstens limitiert die geringe Stichprobengröße und Zahl annotierter Dyskinesie-Episoden im PACMAN-Datensatz die statistische Kraft und begünstigt Varianz in den Modellleistungen. Zweitens ist die gegenwärtige Pipeline auf eine einzelne standardisierte motorische Untersuchung (UPDRS 3.6) fokussiert. Es bleibt offen, inwieweit andere klinisch relevante motorische Prüfungen in gleicher Weise semantisch projizierbar sind. Drittens wurden ausschließlich Bewegungsdaten ausgewertet und viertens ist unklar, ob semantische Reduktionsverfahren bei breiterer Variation von Alltagskontexten strukturell an Gültigkeit verlieren. Fünftens ergeben sich weiterführende

Forschungsfragen. Insbesondere ist zu prüfen, ob eine multimodale Erweiterung um zusätzliche Sensorkanäle (etwa Gyroskopsensorik oder Elektromyographie) die Mustererkennung verbessert, ohne die Interpretierbarkeit der Modelle zu beeinträchtigen.

Gleichzeitig eröffnen sich aus den Befunden mehrere Perspektiven. Die konsistent nachgewiesene Generalisierungsfähigkeit semantisch oder hybrid repräsentierter Bewegungsdaten unterstreicht das Potenzial, solche Ansätze als Bestandteil klinischer Entscheidungsunterstützung einzusetzen, etwa zur engmaschigeren Titrationssteuerung dopaminerger Medikation. Eine Ausweitung der Datengrundlage um größere, multizentrische und stärker diversifizierte Kohorten erscheint erforderlich, um belastbare Konfidenzen und Subgruppenanalysen zu ermöglichen. Ergänzend könnten synthetisch generierte oder augmentierte Zeitreihen zur effizienten Erweiterung datenknapper Phänotypen beitragen. Ferner bleibt die Standardisierung von Erhebungsprotokollen, Metriken und neurologischen Bewertungsgrundlagen zentral für zukünftige Vergleichbarkeit und Modelltransfer.

Trotz der skizzierten Limitationen zeigen die Ergebnisse, dass semantisch abgestützte Merkmalsräume eine tragfähige Grundlage für interpretierbare und klinisch anschlussfähige Klassifikatoren bilden können und damit einen Ansatzpunkt für die Entwicklung robuster, kontextsensitiver Entscheidungssysteme in der Parkinson-Versorgung liefern.

1.5 Fazit

Zusammenfassend demonstriert diese Dissertation systematisch die methodische Überlegenheit semantisch fundierter Merkmalsrepräsentationen und hybrider Ansätze gegenüber rein datengetriebenen Verfahren in Bezug auf klinische Generalisierbarkeit und Interpretierbarkeit. Sie bestätigt damit den Wert klinisch verankerter Vorverarbeitungsschritte, insbesondere in datenlimitierten oder heterogenen Anwendungen. Für eine breite klinische Implementierung sind jedoch multizentrische Kohorten und standardisierte Labelprotokolle notwendig, um die Modellpräzision weiter zu stabilisieren. Insgesamt bietet die Studie eine solide methodische Grundlage für künftige sensorbasierte Erkennungssysteme in der Parkinsondiagnostik.

An diese Arbeit knüpfen mittlerweile zwei weitere klinisch orientierte Studien an, die jeweils rund 100 Parkinson-Patient:innen umfassen. In Übereinstimmung mit den hier gewonnenen Erkenntnissen werden relevante UPDRS-Bewertungen engmaschig durch Mediziner:innen durchgeführt und mit Sensorsystemen synchronisiert. Ziel dieser Folgeprojekte ist es, mithilfe vernetzter Systeme eine möglichst objektive Darstellung des aktuellen Krankheitszustands zu erreichen, um perspektivisch eine algorithmisch gestützte Therapieoptimierung innerhalb der PKB zu ermöglichen. Die in dieser Dissertation beschriebenen Methoden und Ergebnisse leisten damit einen unmittelbaren Beitrag zur Weiterentwicklung sensorbasierter Verfahren zur Erkennung motorischer Symptome bei Parkinson.

2 Artikel im Original

sensors

MDPI

Article

Opportunities and Limitations of Wrist-Worn Devices for Dyskinesia Detection in Parkinson's Disease

Alexander Johannes Wiederhold [1,*], Qi Rui Zhu [1], Sören Spiegel [1], Adrin Dadkhah [2,3], Monika Pötter-Nerger [4], Claudia Langebrake [2], Frank Ückert [1] and Christopher Gundler [1]

[1] Institute for Applied Medical Informatics, University Medical Center Hamburg-Eppendorf, Martinistraße 52, 20246 Hamburg, Germany
[2] Hospital Pharmacy, University Medical Center Hamburg-Eppendorf, Martinistraße 52, 20246 Hamburg, Germany
[3] Department of Stem Cell Transplantation, University Medical Center Hamburg-Eppendorf, Martinistraße 52, 20246 Hamburg, Germany
[4] Institute of Neurology, University Medical Center Hamburg-Eppendorf, Martinistraße 52, 20246 Hamburg, Germany
* Correspondence: a.wiederhold@uke.de; Tel.: +49-176-70073562

Highlights

- Sensor-driven measurements can support the assessment of dyskinesia fluctuations in clinical practice.
- Dimensional reduction of accelerometer data using PCA improves machine learning model performance.
- Semantic feature extraction enhances model generalization and predictive capability for dyskinesia detection.
- Integrating standardized neurological assessments can further improve the reliability of sensor-based monitoring.

Abstract

During the in-hospital optimization of dopaminergic dosage for Parkinson's disease, drug-induced dyskinesias emerge as a common side effect. Wrist-worn devices present a substantial opportunity for continuous movement recording and the supportive identification of these dyskinesias. To bridge the gap between dyskinesia assessment and machine learning-enabled detection, the recorded information requires meaningful data representations. This study evaluates and compares two distinct representations of sensor data: a task-dependent, semantically grounded approach and automatically extracted large-scale time-series features. Each representation was assessed on public datasets to identify the best-performing machine learning model and subsequently applied to our own collected dataset to assess generalizability. Data representations incorporating semantic knowledge demonstrated comparable or superior performance to reported works, with peak F_1 scores of 0.68. Generalization to our own dataset from clinical practice resulted in an observed F_1 score of 0.53 using both setups. These results highlight the potential of semantic movement data analysis for dyskinesia detection. Dimensionality reduction in accelerometer-based movement data positively impacts performance, and models trained with semantically obtained features avoid overfitting. Expanding cohorts with standardized neurological assessments labeled by medical experts is essential for further improvements.

check for updates

Academic Editor: Lorenzo Scalise

Received: 16 June 2025
Revised: 12 July 2025
Accepted: 17 July 2025
Published: 21 July 2025

Citation: Wiederhold, A.J.; Zhu, Q.R.; Spiegel, S.; Dadkhah, A.; Pötter-Nerger, M.; Langebrake, C.; Ückert, F.; Gundler, C. Opportunities and Limitations of Wrist-Worn Devices for Dyskinesia Detection in Parkinson's Disease. *Sensors* **2025**, *25*, 4514. https://doi.org/10.3390/s25144514

Keywords: Parkinson's disease; dyskinesia; wearable devices; semantic data analysis; tsfresh; machine learning; accelerometer data; supportive decision making; feature extraction; principal component analysis

1. Introduction

The optimization and fine-tuning of therapy in response to the progression of Parkinson's disease (PD) necessitate in-hospital diagnostic assessments and evaluations by means of clinical rating scales [1,2]. The overdosing of dopaminergic drugs often evokes levodopa-induced dyskinesia (LID), an uncomfortable side effect that is characterized by uncontrolled, involuntary muscle movements of all body parts [3]. While many symptoms associated with PD may be more effectively identified through observational means, dyskinesias in the upper limbs hold significant potential for detection using wrist-worn sensors. The direct recording of health parameters provides vital insights into the ongoing progression of motor symptoms. Unlike a clinician who can only assess a snapshot of a moment, body-mounted systems are worn continuously throughout the day, offering a temporal recording of events. The acquired data is confined to the specific body part where the sensor is worn, yet it is tailored precisely to its designated tasks, thereby facilitating the capture of therapy-dependent motor fluctuations [4,5]. Subsequently, wearable devices such as smartwatches provide accessible monitoring aids, enabling the recording and tracking of movement data for understanding motor symptom variations and optimizing therapy plans [6–8].

Wearables offer potential benefits for observing daily fluctuations in a hospital setting, allowing an expert-guided labeling of symptoms. While various sensor-driven approaches for capturing tremor and bradykinesia have been explored, there is a limited focus on dyskinesia detection despite its potential as a key biomarker for therapy responses [9,10]. Alongside other reports, studies conducted by Hssayeni et al. and Pfister et al. aim to assess dyskinesia in a free-living environment to be able to monitor the disease condition at home [11,12]. Hssayeni et al. seeks to estimate dyskinesia using accelerometer data during daily activities, reporting a Pearson correlation within the range of 0.70 to 0.84. In a similar vein, Pfister et al. reports the capability to detect dyskinesias in a free-living environment, achieving a sensitivity/specificity of 0.64/0.89. Acknowledging the importance of understanding the symptoms impact on everyday activities, there is clinical relevance in observing and estimating its occurrence during a hospital stay. When patients visit the clinic, neurologists adjust dopaminergic pharmaceuticals, leading to the frequent appearance of LID [3,13]. Monitoring these side effects during hospital admission could provide insights into the symptomatic fluctuations of patients, potentially aiding in the determination of the optimal drug dosage for individuals and supporting clinicians in addressing LIDs before patients are discharged. Sieberts et al. conducted a study referred to as the DREAM Challenge, which aligns with this suggestion by identifying biomarkers linked to tremor, bradykinesia, and dyskinesia in PD. Utilizing public datasets, the DREAM Challenge aimed to predict the severity of PD symptoms, resulting in an AUPR (area under the precision-recall curve) of 0.48 specifically for dyskinesia [9].

Subsequently, we encourage the implementation of a novel monitoring setup specifically designed for dyskinesias emerging during hospital admission. Additionally, we propose an evaluation of the generalizability of movement data from publicly accessible datasets to our internally collected hospital data.

Thus, this paper employs two innovative approaches for movement data representation: one is a purely semantic technique utilizing principal component analysis (PCA) in

combination with biomechanical feature extraction, and the other is an automatic, serial feature representation. Following the training of these methods on publicly available datasets from the Michael J. Fox Foundation (MJFF) [14], our objective was to assess the performance of the resulting models on our own collected movement data and thus the models' generalizability. To fulfill our objective, we formulated two intents: (1) investigating the impact of various movement data representations on model performance and (2) evaluating the generalizability of machine learning (ML) models from publicly available datasets [14] onto the PACMAN (Parkinson's Clinical Movement Assessment) dataset.

2. Materials and Methods

2.1. Data

2.1.1. Public Datasets

We used two distinct and publicly available datasets from the MJFF: the Levodopa Response Study (LRS) and the Clinician Input Study (CIS-PD) [14,15]. Both studies aim to measure movement symptoms and their fluctuations of PD by means of accelerometers and according to the Unified Parkinson's Disease Rating Scale (UPDRS). These datasets were always retrieved together and referred to as MJFF dataset.

The LRS includes 28 patients diagnosed with PD that were monitored both in-clinic, where they engaged in a battery of standard activities, and at home while performing their daily activities. While the primary focus of the study is to comprehend motor fluctuations, patients were measured over four consecutive days with accelerometers. On the day of admission, UPDRS assessments were conducted at the clinic while patients were still on regular dopaminergic medication. Over the next two days, patients were released home, where they could carry out regular activities. On the last day, patients returned to the clinic and underwent similar UPDRS tests without any dopaminergic treatment [14].

The CIS-PD was a 6-month longitudinal investigation involving wearable tracking for 51 PD patients. The study encompassed clinic visits and at-home monitoring using smartwatches. Following the baseline assessment, in-clinic visits were scheduled at 2 weeks, 1 month, 3 months, and 6 months. During these visits, clinicians conducted standard clinical assessments and reviewed data recorded at home. Between the hospital visits, the patients were asked to continue wearing their smartwatches and regularly report symptom severity and medication intake using a mobile phone app [15].

2.1.2. Data Collection of Our Own Movement Data (PACMAN)

PD patients admitted at the Department of Neurology at the University Medical Center Hamburg-Eppendorf (UKE) stay a minimum of two weeks during an inpatient care program, known as the Parkinson-Komplexbehandlung (PKB). A multidisciplinary team, including neurologists, physiotherapists, neuropsychologists, and other paramedical specialists, is dedicated to a patient-centered and individualized clinical approach in search of the optimal therapy [2,15]. This setting presents a distinctive opportunity for the continuous gathering of accelerometer data, accompanied by task-coupled severity scoring.

During our own 4-month clinical data collection, we raised 7 different standardized clinical examinations according to the third part of the Unified Parkinson Disease Rating Score (UPDRS III) per visit. An experienced neurologist decided on the examination criteria relevant for our hypothesis to detect dyskinesias during hospital admittance. The neurological task was required to not only be significant for the detection of upper limb dyskinesia but also measurable by a wrist-worn device. Hence, we decided on alternating hand movements, which include the supination and consequent pronation of the most affected hand (UPDRS 3.6). Further, we assumed that this periodic movement is projectable by semantic data representation and thus offers potential for clinically relevant feature extraction.

Each patient underwent a minimum of two daily visits over a span of up to two weeks, mirroring the duration of the PKB program [2]. During the initial consultation, the physician provided each patient with the watches and conducted the necessary setup each morning. Subsequently, the UPDRS 3.6 task was assessed, and the corresponding timestamp was annotated for accurate measurement. To ensure consistent data quality, the physician verified that the watch was worn tightly and in the correct orientation during each labeled assessment. The watch remained on the patient's most affected side, continuously capturing accelerometer data until the physician's return in the afternoon.

This non-interventional prospective cohort study used the accelerometer of an Apple Watch Series 6. The resulting dataset, further referred to as PACMAN, comprises movement data along with timestamped labels indicating symptom severity. The entire data collection transpired within the framework of clinical routine at an inpatient unit of the UKE and was carried out in accordance with relevant guidelines and regulations (Declaration of Helsinki). Written informed consent for the study was obtained from all participants and an approval from the Ethics Commission of the Ärztekammer Hamburg under the ID 2022-100846-BO-ff was granted beforehand.

2.2. Uniform Data Infrastructure

To achieve our goal of robustly detecting motor fluctuations, it was imperative to establish an infrastructure that could seamlessly integrate into clinical settings. This need was met by implementing a database adhering to the Fast Healthcare Interoperability Resources (FHIR) standard, ensuring the consistent storage and retrieval of movement data [16]. All acquired movement data is systematically deposited into the FHIR database, serving as an objective for subsequent comparative analyses. Once stored, a custom-designed data loader facilitates the retrieval of all measurements, allowing for tailored specifications and consistent loading of the requested data. Our team devised this uniform data architecture in a preceding project, with comprehensive details published elsewhere [16].

The process of obtaining all stored movement data involves a crucial consideration regarding the length of measurement samples. In the realm of sequential health data analysis, the term window size refers to the duration in which the data is examined. Given that different measurements possess unique recording lengths, the sample's window size can be of standardized length or dynamic. Opting for a predefined length ensures straightforward cross-database compatibility but may result in an exclusion of shorter measurements. Thus, setting the window size to a smaller value may omit a necessary task characteristic in some samples [17]. Therefore, we retain the full duration of each task as observed in the clinical setup and store them as variable-length samples. All subsequently chosen methods were selected to be compatible with this kind of temporal data with varying lengths. To further ensure comparability across sessions and participants, we applied a global rotation to compensate for the different coordinate systems of the chosen wearables. Since each sample was labeled in its entirety by a clinician, it is important to preserve the full temporal context. Subsequent classification directly relies on these clinical labels and thus benefits from maintaining the integrity of the original task as far as possible.

2.3. Data Representations

Representing movement data is crucial for its adequate analysis. A data representation is a way of encoding information into a format interpretable by an algorithm without losing significant meaning. Therefore, the selected representation format should align with the inherent characteristics of the measured data. Employing an appropriate data representation facilitates the extraction of meaningful features during preprocessing. The resulting features are, in turn, used to train our models for the detection of dyskinesias. We

opted for two distinct techniques of data representation that maintain the structure of the acquired accelerometer recordings.

2.3.1. Semantic Representation: PCA and Biomechanical Features

As movement data is recorded in a three-dimensional space by accelerometers, a reduction in dimensionality is amenable to visualization and analysis through human interaction. Given our focus on the alternating hand movements, we assumed that this periodic movement should be prominent in the data. This unique characteristic provides an opportunity for the representation of multi-dimensional movement data using semantic approaches. To reduce the complexity of our data, while still respecting the activity on each axis, we utilized a PCA.

PCA was applied to project the tri-axial accelerometer data into a single principal component, yielding a one-dimensional representation of movement over time. This projection facilitates semantic feature extraction, where distinct movement characteristics can be interpreted in a way that mirrors clinical assessment. As depicted in Figure 1, this allows clinicians and researchers alike to identify meaningful signal patterns, such as oscillation symmetry or amplitude modulation, using a reduced yet expressive representation.

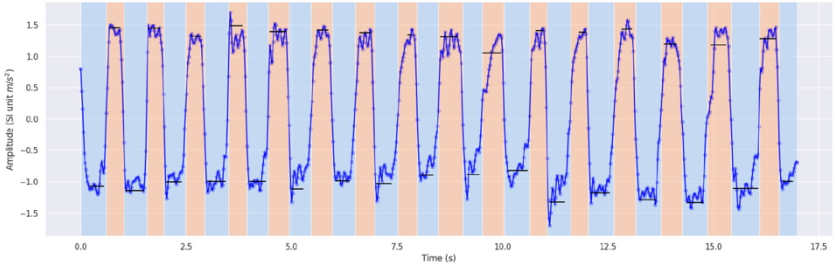

Figure 1. An exemplary segmented measurement. The blue segments have their extremum smaller than 0 and the orange segments show an extremum larger than 0. This segmentation is used for feature analysis.

Our semantic feature extraction follows the methodology introduced by Sánchez-Fernández et al., who identified 20 biomechanical features as particularly relevant for characterizing the alternating hand movement task (UPDRS 3.6) [18]. Their analysis, however, relied on multimodal sensor input (accelerometer, gyroscope, magnetometer). Since our dataset comprises only accelerometry data, we selected a subset of 9 clinically and technically interpretable features that can be derived robustly from this modality. These were selected through expert-driven, domain-specific relevance. The selected features are as follows:

- Feature for entire measurement:
  - Absolute mean of extremum
- Features for each part:
  - Number of segments comprising each part
  - Duration ratio of each part
  - Mean duration of each part
  - Interquartile range of each part
  - Relative maximum mean of each part
  - Relative maximum interquartile range of each part
  - Relative minimum mean of each part
  - Relative minimum interquartile range of each part

Next, to respect characteristic events of each individual measurement, we divided each sample into three distinct parts for the beginning, middle, and end. These parts are further divided into equally enduring segments. Figure 1 shows an example measurement with segments marked in orange or blue, defined by either the segment's maxima or minima, respectively. This segmentation was crucial for capturing temporal variations within the movement, such as oscillatory or acceleration changes, which are diagnostically relevant in Parkinson's disease.

2.3.2. Automatic Feature Extraction

In our alternative representation method, we utilized an automatic time-series feature extraction tool, further referred to as tsfresh, to automate the complex process of time-series engineering. Instead of human guidance, the Python library (Version: 0.20.1 on Python 3) identifies features by considering different algorithms of signal processing and time-series analysis to extract over 3000 features from temporal structured data [19]. This vast amount of mostly interpretable features is then systematically reduced through statistical tests.

As the automatically feature extraction can be applied on both the original three-dimensional as well as the reduced data, we considered three ways of movement data representation: (1) a fully semantic technique consisting of PCA and biomechanical feature extraction, (2) the PCA combined with the automatic feature extraction, and (3) automatic feature extraction only.

2.4. Training of the ML Models

Each of the three representation techniques yields distinctive features for the measured UPDRS task. A vast amount of these features, especially the numerous automatically extracted features, are not relevant to our research question. However, to identify an optimal ML model that maximizes performance based on only meaningful features, we aimed to determine the best combination of each representation with available models. The core component of the resulting data pipeline is the classifier, a predefined algorithm that assigns labels based on the provided features. To identify the classifier's highest performance, we integrated every possible combination into a grid search that predicts on the MJFF dataset.

A grid search is a systemic hyperparameter optimization that finds the optimal configuration of meaningful features, the classifier, and its hyperparameters by training each model separately [20]. Our implemented grid search used a stratified 10-fold cross validation prepared for imbalanced data [21]. We considered (1) a feature selector to find the ideal count of relevant features, (2) an oversampling technique to equalize for underrepresented labels (SMOTE) [22], and (3) different classifiers with numerate possibilities of their hyperparameters. As for the selection of parametric and non-parametric classification algorithms, we compared the performance of a logistic regression, a k-nearest neighbors' classifier, a random forest classifier, a support vector machine, and a gradient-boosting classifier.

2.5. Evaluation of the Resulting ML Models

Next, we determined the ten best-performing combinations of classifiers, selectors, and samplers for each of the three methods of feature representation on the MJFF dataset. The grid search results were ranked by the unweighted F_1 score, calculated as the arithmetic mean of all per-class F_1 scores. By combining precision and recall into a single metric, the F_1 score offers a balanced view of the model's performance across both classes. Although no consensus exists on metric selection for clinical relevance in this particular task [11,12] a high F_1 score suggests greater reliability in capturing label fluctuations, a critical requirement for potential therapeutic decision support. In addition to the F_1 score, the accuracy of each model's performance was also computed.

To ensure robustness of our findings, we conducted a conventional t-test, incorporating Welch's modification to accommodate potential variations and to assess significant performance disparities among different models. We expressed our outcomes as mean ± standard deviation, with a predefined threshold for statistical significance set at $p < 0.001$.

As a final step, we employed the top 10 models per representation, based on their unweighted F_1 performance, and implemented them on our own collected PACMAN movement data. The assessment of their performance utilized the same metric analysis and statistical significance to ensure comparability.

3. Results

3.1. The Patient Cohort

Our paper incorporated a total of 27 patients from the LRS dataset, spanning an age range of 50 to 84 years, with an average age of 67 years (±9 years). We included an average of 51 dyskinesia measurements (±9 measurements) per patient. Additionally, we included 24 patients from CIS-PD, with an average age of 63 years (±10 years), ranging from 36 to 75 years. Here, we used 12 dyskinesia measurements (±3 measurements) per patient on average. Finally, our in-clinic data collection contributed 25 patients, with an average age of 65 years (±8 years) and a range from 49 to 84 years. The PACMAN dataset incorporates an average of 3 dyskinesia measurements (±2 measurements) per patient.

3.2. Data Integrity

The presented data sources originate from distinct sites and were designed for different purposes. Nevertheless, they share comparability in terms of sensor type, demographics, neurological assessments, and disease-related intention of recording. The populations depicted in all MJFF studies fall within the same age range and undergo recurring hospital care for therapy adjustments. The types of sensors employed are consistent, as all studies integrate accelerometers, with both CIS-PD and PACMAN even utilizing Apple Watch devices.

Given that the MJFF studies involve a considerable number of ambulatory accelerometer recordings without precise physician annotations, our exclusive reliance on supervised labels was necessary to achieve our goal of identifying the most accurate representation of movement data for dyskinesia detection in a clinical setting. Consequently, all retained movement data and its corresponding labels are derived from hospital admittance and adhere to UPDRS standards.

3.3. Performance on Training Datasets

Figure 2 illustrates the confusion matrix of the leading model for each data representation on the MJFF studies. This visualization simultaneously presents the assigned dyskinesia label and the model's prediction. Hence, the confusion matrix provides a reliable means to identify instances when a model incorrectly labeled a class. All depicted models adequately identified the absence of dyskinesia. However, they encountered challenges in accurately recognizing dyskinesia.

When it comes to the full evaluation of performance, we ranked the predictive outcomes for each representation by the unweighted F_1 score. Here, we achieved peak performances of 0.68 with both data representations, employing solely automatically extracted features and the automatically extracted features on the one-dimensional representation. The semantic representation using PCA and biomechanical features yielded a slightly lower performance, registering 0.63 for unweighted F_1. Accuracies of all three extraction methods

were as high as 0.89 for each representation. The top five models per representation are listed in Tables 1–3.

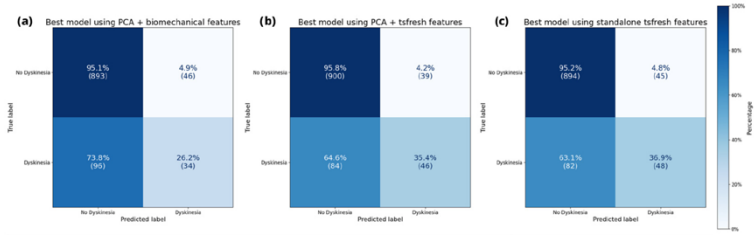

Figure 2. Confusion matrix of the best-performing models for each representation method on the MJFF dataset.

Table 1. Results of top 5 models for representation: PCA and biomechanical features on the MJFF dataset.

Rank	Model			Macro F_1 Score	Accuracy
	Classifier	Selector	Sampling		
1	gradient boosted trees (lr = 1.0)	10	none	0.63	0.87
2	gradient boosted trees (lr = 1.0)	none	none	0.62	0.88
3	random forest	5	SMOTE	0.61	0.89
4	gradient boosted trees (lr = 1.0)	none	SMOTE	0.60	0.83
5	k-nearest neighbors (nn = 5)	5	none	0.60	0.88

lr = learning rate, nn = number of neighbors, SMOTE = Synthetic Minority Over-sampling Technique.

Table 2. Results of top 5 models for representation: PCA and tsfresh features on the MJFF dataset.

Rank	Model			Macro F_1 Score	Accuracy
	Classifier	Selector	Sampling		
1	random forest	none	SMOTE	0.68	0.88
2	random forest	10	none	0.68	0.88
3	random forest (depth = 5)	10	none	0.67	0.89
4	random forest (depth = 5)	none	SMOTE	0.67	0.84
5	gradient boosted trees (lr = 1.0)	10	none	0.66	0.86

lr = learning rate, SMOTE = Synthetic Minority Over-sampling Technique.

Table 3. Results of top 5 models for representation: PCA and standalone tsfresh features on the MJFF dataset.

Rank	Model			Macro F_1 Score	Accuracy
	Classifier	Selector	Sampling		
1	random forest	10	none	0.68	0.88
2	random forest (depth = 5)	10	none	0.67	0.89
3	random forest (depth = none)	none	SMOTE	0.67	0.88
4	random forest (depth = 5)	none	SMOTE	0.67	0.83
5	random forest (depth = 5)	5	none	0.66	0.88

SMOTE = Synthetic Minority Over-sampling Technique.

Neither the purely semantic technique, nor the automatic feature extraction, alone or combined, achieved a significant difference compared to each other. Across all performed combinations, the average unweighted F_1 yielded scores of 0.53 ± 0.05, 0.57 ± 0.08, and 0.57 ± 0.08 for PCA with biomechanical features, PCA with feature extraction, and standalone feature extraction, respectively (mean $\pm$ standard deviation).

3.4. Generalization into Clinical Setting

While the performance on the public datasets though cross validations might approximate the generalizability on unseen data, we further tested this claim by utilizing the novel clinical data collected for this study. Application of the best 10 models per representation method on the PACMAN validation set yielded nuanced findings. The confusion matrix of the best model per representation, depicted in Figure 3, shows a similar classification pattern as on the MJFF studies.

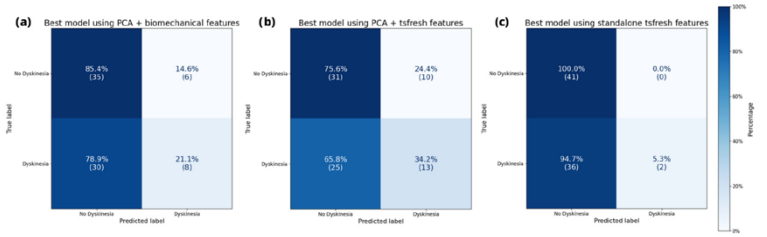

Figure 3. Confusion matrix of the best-performing models for each representation on the PACMAN dataset.

The top-performing representation for unweighted F_1 achieved a score of 0.53, utilizing PCA in conjunction with automatically extracted features. Following this was the semantic approach with a score of 0.48, and the standalone automatically extracted features ranked the lowest with a score of 0.40. The top five outcomes per representation of our generalization efforts are detailed in Tables 4–6, including all the parameters employed.

Table 4. Results of top 5 models for representation: PCA and biomechanical features on PACMAN dataset.

Rank	Model			Macro F_1 Score	Accuracy
	Classifier	Selector	Sampling		
1	gradient boosted trees (lr = 1.0)	none	SMOTE	0.48	0.54
2	gradient boosted trees (lr = 1.0)	10	SMOTE	0.47	0.52
3	random forest	10	SMOTE	0.45	0.53
4	gradient boosted trees (lr = 1.0)	5	none	0.44	0.53
5	gradient boosted trees (lr = 1.0)	10	none	0.41	0.51

lr = learning rate, SMOTE = Synthetic Minority Over-sampling Technique.

Table 5. Results of top 5 models for representation: PCA and tsfresh features on PACMAN dataset.

Rank	Model			Macro F_1 Score	Accuracy
	Classifier	Selector	Sampling		
1	gradient boosted trees (lr = 1.0)	10	none	0.53	0.56
2	gradient boosted trees (lr = 1.0)	none	none	0.44	0.56
3	random forest	5	none	0.42	0.54
4	random forest (depth = 5)	5	none	0.40	0.54
5	random forest (depth = 5)	none	SMOTE	0.38	0.51

lr = learning rate, SMOTE = Synthetic Minority Over-sampling Technique.

Overall, the three methods yielded average unweighted F_1 scores of 0.42 ± 0.04, 0.39 ± 0.06, and 0.36 ± 0.02 for PCA with biomechanical features, PCA with automatically extracted features, and automatically extracted features alone, respectively (mean $\pm$ standard deviation). Performances of the purely semantic representation were significantly higher than the automated technique ($p < 0.001$).

Table 6. Results of top 5 models for representation: standalone tsfresh features on PACMAN dataset.

Rank	Model			Macro F_1 Score	Accuracy
	Classifier	Selector	Sampling		
1	k-nearest neighbors (nn = 5)	10	none	0.37	0.54
2	random forest	10	none	0.37	0.53
3	random forest (depth = 5)	5	none	0.37	0.53
4	gradient boosted trees (lr = 1.0)	10	none	0.37	0.49
5	random forest (depth = 5)	none	SMOTE	0.36	0.52

lr = learning rate, nn = number of neighbors, SMOTE = Synthetic Minority Over-sampling Technique.

4. Discussion

This paper explores ML models for movement data representation, utilizing two distinct approaches and their combination. We first determined the top-performing models on the MJFF datasets and then evaluated their performance on our collected test dataset. Thereby, we were focusing on the impact of movement data representations and the generalizability of our resulting models for dyskinesia detection as our central hypotheses.

The results demonstrate how a dimensional reduction in movement data informed by the nature of the task has a positive impact on performance. Irrespective of whether combined with semantic or automatic extraction methods, the top 10 performing models incorporate this transformation when applied on our PACMAN dataset. The optimal performance is observed when the transformation is combined with automatic features. Nevertheless, the transformation combined with biomechanical features yielded comparable performance. This finding supports the idea that human-interpretable features enhance the ability of ML techniques to generalize across movement datasets. This finding confirms that semantically grounded preprocessing can serve as a safeguard against overfitting while supporting interpretability, a crucial factor in clinical implementation. The models using unselected features from a single automatic feature extraction generally show better results during training on the MJFF datasets but fail on our collected PACMAN data. Likely due to overfitting, it almost entirely fails to detect dyskinesia labels and assigns only two labels correctly. This underlines the importance of aligning feature representations with domain knowledge to improve robustness in real-world deployment.

Comparing the two approaches in detail, we observe that automated feature extraction (tsfresh) offers a wide array of statistical descriptors that may capture subtle signal characteristics, which contributes to strong performance on the structured and relatively homogeneous MJFF dataset. However, this approach appears to be less robust when applied to the clinically diverse PACMAN dataset, suggesting high sensitivity to variability in sensor noise, wearing conditions, and patient behavior. In contrast, the semantic representation consistently yields more stable results, even with limited and heterogeneous data. This suggests that the semantic approach not only improves interpretability for clinicians but also enhances robustness against real-world variability, which is critical for generalization across datasets. Therefore, while automated features may excel in high-data or controlled settings, semantically grounded representations prove more effective in noisy, low-data clinical environments, where reliability and explainability are essential.

Further, our analysis aligns closely with the performance reported in the DREAM Challenge, which aimed to identify LIDs on MJFF datasets resulting in an AUPR of 0.48. However, the latter study did not evaluate distinct data representations nor test generalization on an independent dataset [9]. Previously, the mentioned studies by Hssayeni et al. and Pfister et al. report higher performances in dyskinesia detection, but their non-clinical setups are incomparable in terms of the sensor types used or a non-standardized assessment of dyskinesia [11,12]. Moreover, almost all the presented papers reveal distinct

metrics, which are incomparable to each other. As each metric analysis is favorable for the individual intention, these metrics are a challenge to evaluate in terms of comparability. To this end, we chose the F_1 score as the principal metric for performance evaluation, as it balances precision and recall, two properties particularly important in the clinical context where both false positives and false negatives can impact therapeutic decisions.

While there is currently no universally accepted threshold for the F_1 score in clinical ML applications, our observed values (MJFF: 0.68; PACMAN: 0.53) indicate a level of consistency and reliability that supports potential real-world use.

In order to lay the groundwork for ML generalization to work, standardization is required. First, the evaluative framework of studies working with supervised ML on movement data should use comparable and clinically relevant metrics. Analyses of medical data must account for class-specific performance, as predictions for each class need to be evaluated separately. Overall accuracy, for example, is insufficient in clinical settings, as it can obscure the detection of relevant disease phenotypes, particularly in imbalanced datasets. Secondly, the application of standardized neurological assessments, such as the suggested UPDRS, should be used. While some publications evaluate activities of daily living [11], these activities are not sufficiently reproducible and only play a minor role in therapeutic adjustments. These assessments lay the foundation of the task-specific data labeling and thus are essential for generalizability.

On the contrary, the data collection approach presented in this paper provides a unique opportunity to acquire standardized clinical assessments of movement distortions over a two-week period per patient. The dense data quality obtained per patient facilitates the detection of dyskinesias in a hospital setting enabling early identification of LIDs for medication adjustments before the patient is discharged.

Regarding clinical implementation of the presented movement data methodology, the generalizability of the fully semantic representation suggests great opportunities for future applications of wearables to detect LIDs during clinical stays. Our primary assumption, that the periodic alternation between supination and pronation of the hand are well projectable by a simple dimensional reduction, turns out to be valid. Adhering to clinical expertise and translating it directly into straightforward data representations suitable for machine learning algorithms significantly influences the final performance outcomes. Although real-time analysis was not the focus of this study, the short inference time of our trained models suggests feasibility for future real-time applications, such as adaptive therapy monitoring during inpatient stays. The dimensionally reduced representations demonstrated better generalization on the PACMAN dataset compared to the automatic features, which overfitted on the MJFF dataset. This supports the value of embedding domain knowledge into preprocessing pipelines, particularly when data availability is limited, a common challenge in real-world clinical contexts. The results on the PACMAN dataset further suggest the enormous potential for a combination of both techniques. Combining task-specific semantic dimensional reduction with automatic feature extraction may offer the best of both worlds, as this hybrid approach performed best on the PACMAN dataset. Both techniques are rooted in the assumption that the UPDRS examination, a clinically validated neurological scoring standard for over 30 years, provides a robust foundation for interpreting dyskinesia. Accordingly, the semantic pipeline partially mimics a clinician's process of evaluating movement patterns. A multidisciplinary approach between clinical expertise and data science is imperative for a successful application of this technology into routine.

However, this technology has its limitations for the detection of dyskinesia due to its unreliable predictive power. The relatively small size of our PACMAN dataset constrains the statistical power of our findings and likely contributes to variability in model

performance. This study was designed as an empirical step towards estimating minimal data requirements under clinical constraints, but future work must include larger, statistically powered datasets. Particularly, an expanded data collection of clinically annotated measurements is essential for tracking LIDs and assisting clinicians in optimizing therapy. Perhaps even synthetically generated movement data could provide a foundation to train and optimize models without the extensive need of gathering patients. Furthermore, comparability between datasets and standardization plays a pivotal role for the generalizability and its coherent ability to detect dyskinesias, as stated previously. Also, it remains unclear if other neurological assessments that are diagnostically relevant for the dyskinesia detection can be projected by semantic knowledge. This suggestion is also undermined by the smartwatch's limitation to detect movement of fingers or the hand. Nevertheless, the outcomes of this study demonstrate potential for the development of robust decision-support systems grounded in semantic principles, particularly in the analysis of clinically recorded movement data.

5. Conclusions

This paper presents a unique opportunity to gather standardized clinical assessments of movement distortions over a two-week period per patient, facilitating enhanced dyskinesia detection within a hospital setting. The utilization of inbuilt accelerometers in smart watches provides a stable and convenient solution to track movement distortions in hospitalized patients. Our investigation has identified objective movement data representations conducive to dyskinesia recognition and highlighted the semantic impact on sensor data analysis, especially when generalizing onto foreign datasets. Notably, semantic feature representations demonstrated more robust performance than automated features when applied to real-world clinical data. This robustness stems from their interpretability and alignment with established clinical rating schemes. The results suggest that semantically grounded preprocessing may offer a critical advantage in small-data or high-variability scenarios. Yet, further research with a larger cohort and standardized labeling protocols is essential to optimize therapy for levodopa-induced dyskinesias. The results of this work provide evidence of feasibility, suggesting that technology-based measurements have the potential to serve as supportive tools for comprehending symptom fluctuations during clinical practice.

Author Contributions: Conceptualization, A.J.W., Q.R.Z., S.S., A.D., M.P.-N., C.L., F.Ü. and C.G.; methodology, A.J.W. and C.G.; software, Q.R.Z., S.S., C.G. and A.J.W.; validation, A.J.W. and C.G.; formal analysis, A.J.W. and C.G.; investigation, A.J.W.; resources, A.J.W.; data curation, A.J.W.; writing—original draft preparation, A.J.W.; writing—review and editing, A.J.W., Q.R.Z., S.S., A.D., M.P.-N., C.L., F.Ü. and C.G.; visualization, A.J.W.; supervision, C.G.; project administration, A.J.W.; funding acquisition, not applicable. All authors have read and agreed to the published version of the manuscript.

Funding: No financial support was received for the preparation of this manuscript. Apple Inc. provided the Apple Watch devices for the research. Apple was not involved in the design of the research, nor was it involved in the collection, analysis, or interpretation of the research data, or the content of this or any related publication. We acknowledge financial support from the Open Access Publication Fund of UKE—Universitätsklinikum Hamburg-Eppendorf.

Institutional Review Board Statement: Participants in the study (PACMAN) granted informed consent. All procedures involving human participants adhered to the ethical standards of the institutional research committee and were in accordance with the 1964 Helsinki Declaration and its subsequent amendments or comparable ethical standards. The collection of patient measurements received approval from the Ethics Commission of the Ärztekammer Hamburg under the ID 2022-100846-BO-ff.

Informed Consent Statement: Written informed consent was obtained from all subjects involved in the study (PACMAN).

Data Availability Statement: The code generated during the current study is publicly available in the repository of the University of Hamburg, (http://doi.org/10.25592/uhhfdm.14189 (accessed on 16 July 2025) since 18 July 2025.

Acknowledgments: The authors extend their appreciation to the Michael J. Fox Foundation for Parkinson's Research for generously providing access to the datasets associated with the Levodopa Response Study and the Clinician Input Study to the scientific community. A.J.W. expresses his gratitude to Sara Tiedemann for her valuable contributions during the manuscript preparation. During the preparation of this work the authors used DeepL Translator and ChatGPT 4o in order to enhance the language and style of specific sections. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Conflicts of Interest: The authors declare that they have no competing interests.

Abbreviations

The following abbreviations are used in this manuscript:

UKE	University Medical Center Hamburg-Eppendorf
PCA	Principal Component Analysis
tsfresh	Time-Series Feature Extractors
LID	Levodopa-Induced Dyskinesia
PD	Parkinson's Disease
MJFF	Michael J. Fox Foundation
UPDRS	Unified Parkinson's Disease Rating Scale
LRS	Levodopa Response Study
CIS-PD	Clinician Input Study-Parkinson's Disease
PKB	Parkinson-Komplexbehandlung
PACMAN	Parkinson's Clinical Movement Assessment
FHIR	Fast Healthcare Interoperability Resources
ML	Machine Learning
SMOTE	Synthetic Minority Over-Sampling Technique
AUPR	Area Under the Precision-Recall Curve
DREAM	Digital Biomarker Evaluation and Analysis for Mobile Health Challenge

References

1. Post, B.; Merkus, M.P.; de Bie, R.M.; de Haan, R.J.; Speelman, J.D. Unified Parkinson's disease rating scale motor examination: Are ratings of nurses, residents in neurology, and movement disorders specialists interchangeable? *Mov. Disord.* **2005**, *20*, 1577–1584. [CrossRef]
2. Richter, D.; Bartig, D.; Muhlack, S.; Hartelt, E.; Scherbaum, R.; Katsanos, A.H.; Müller, T.; Jost, W.; Ebersbach, G.; Gold, R.; et al. Dynamics of Parkinson's Disease Multimodal Complex Treatment in Germany from 2010–2016: Patient Characteristics, Access to Treatment, and Formation of Regional Centers. *Cells* **2019**, *8*, 151. [CrossRef]
3. Kwon, D.K.; Kwatra, M.; Wang, J.; Ko, H.S. Levodopa-Induced Dyskinesia in Parkinson's Disease: Pathogenesis and Emerging Treatment Strategies. *Cells* **2022**, *11*, 3736. [CrossRef]
4. Maetzler, W.; Domingos, J.; Srulijes, K.; Ferreira, J.J.; Bloem, B.R. Quantitative wearable sensors for objective assessment of Parkinson's disease. *Mov. Disord.* **2013**, *28*, 1628–1637. [CrossRef]
5. Jalloul, N. Wearable sensors for the monitoring of movement disorders. *Biomed. J.* **2018**, *41*, 249–253. [CrossRef]
6. Bloem, B.R.; Post, E.; Hall, D.A. An Apple a Day to Keep the Parkinson's Disease Doctor Away? *Ann. Neurol.* **2023**, *93*, 681–685. [CrossRef]
7. Giannakopoulou, K.-M.; Roussaki, I.; Demestichas, K. Internet of Things Technologies and Machine Learning Methods for Parkinson's Disease Diagnosis, Monitoring and Management: A Systematic Review. *Sensors* **2022**, *22*, 1799. [CrossRef] [PubMed]
8. Tabatabaei, S.A.H.; Pedrosa, D.; Eggers, C.; Wullstein, M.; Kleinholdermann, U.; Fischer, P.; Sohrabi, K. Machine Learning Techniques for Parkinson's Disease Detection using Wearables during a Timed-up-and-Go-Test. *Curr. Dir. Biomed. Eng.* **2020**, *6*, 376–379. [CrossRef]

9. Sieberts, S.K.; Schaff, J.; Duda, M.; Pataki, B.Á.; Sun, M.; Snyder, P.; Daneault, J.-F.; Parisi, F.; Costante, G.; Rubin, U.; et al. Crowdsourcing digital health measures to predict Parkinson's disease severity: The Parkinson's Disease Digital Biomarker DREAM Challenge. *npj Digit. Med.* **2021**, *4*, 53. [CrossRef] [PubMed]
10. Van Gerpen, J.A.; Kumar, N.; Bower, J.H.; Weigand, S.; Ahlskog, J.E. Levodopa-Associated Dyskinesia Risk Among Parkinson Disease Patients in Olmsted County, Minnesota, 1976-1990. *Arch. Neurol.* **2006**, *63*, 205–209. [CrossRef]
11. Hssayeni, M.D.; Jimenez-Shahed, J.; Burack, M.A.; Ghoraani, B. Dyskinesia estimation during activities of daily living using wearable motion sensors and deep recurrent networks. *Sci. Rep.* **2021**, *11*, 7865. [CrossRef]
12. Pfister, F.M.J.; Um, T.T.; Pichler, D.C.; Goschenhofer, J.; Abedinpour, K.; Lang, M.; Endo, S.; Ceballos-Baumann, A.O.; Hirche, S.; Bischl, B.; et al. High-Resolution Motor State Detection in Parkinson's Disease Using Convolutional Neural Networks. *Sci. Rep.* **2020**, *10*, 5860. [CrossRef]
13. Hughes, A.J.; Daniel, S.E.; Kilford, L.; Lees, A.J. Accuracy of clinical diagnosis of idiopathic Parkinson's disease: A clinico-pathological study of 100 cases. *J. Neurol. Neurosurg. Psychiatry* **1992**, *55*, 181–184. [CrossRef]
14. Synapse.org. MJFF Levodopa Response Study [Internet]. 2019. Available online: https://www.synapse.org/Synapse:syn20681023/wiki/594678 (accessed on 16 July 2025).
15. Elm, J.J.; Daeschler, M.; Bataille, L.; Schneider, R.; Amara, A.; Espay, A.J.; Afek, M.; Admati, C.; Teklehaimanot, A.; Simuni, T. Feasibility and utility of a clinician dashboard from wearable and mobile application Parkinson's disease data. *npj Digit. Med.* **2019**, *2*, 95. [CrossRef] [PubMed]
16. Gundler, C.; Zhu, Q.R.; Trübe, L.; Dadkhah, A.; Gutowski, T.; Rosch, M.; Langebrake, C.; Nürnberg, S.; Baehr, M.; Ückert, F. A Unified Data Architecture for Assessing Motor Symptoms in Parkinson's Disease. *Stud. Health Technol. Inform.* **2023**, *307*, 22–30. [CrossRef] [PubMed]
17. Banos, O.; Galvez, J.-M.; Damas, M.; Pomares, H.; Rojas, I. Window Size Impact in Human Activity Recognition. *Sensors* **2014**, *14*, 6474–6499. [CrossRef] [PubMed]
18. Sánchez-Fernández, L.P.; Garza-Rodríguez, A.; Sánchez-Pérez, L.A.; Martínez-Hernández, J.M. A Computer Method for Pronation-Supination Assessment in Parkinson's Disease Based on Latent Space Representations of Biomechanical Indicators. *Bioengineering* **2023**, *10*, 588. [CrossRef]
19. Christ, M.; Braun, N.; Neuffer, J.; Kempa-Liehr, A.W. Time Series FeatuRe Extraction on basis of Scalable Hypothesis tests (tsfresh—A Python package). *Neurocomputing* **2018**, *307*, 72–77. [CrossRef]
20. Liashchynskyi, P.; Liashchynskyi, P. Grid Search, Random Search, Genetic Algorithm: A Big Comparison for NAS [Internet]. *arXiv* **2019**, arXiv:1912.06059. Available online: http://arxiv.org/abs/1912.06059 (accessed on 27 March 2024).
21. Lemaître, G.; Nogueira, F.; Aridas, C.K. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *J. Mach. Learn. Res.* **2017**, *18*, 559–563.
22. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic Minority Over-sampling Technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]

3 Zusammenfassung

Deutsche Zusammenfassung

Diese Dissertation untersucht die methodischen Grundlagen, Chancen und Limitationen handgelenkbasierter Sensorsysteme zur automatisierten Erkennung von Levodopa-induzierten Dyskinesien bei Parkinson-Patient:innen. Im Zentrum steht die Hypothese, dass semantische Merkmalsräume, orientiert an ärztlicher Beobachtungslogik, eine bessere Generalisierbarkeit auf klinisch fremde Daten ermöglichen als rein datengetriebene Verfahren.

Die Grundlage bildeten zwei öffentliche Datensätze sowie ein eigenständig erhobener klinischer Datensatz mit standardisierten klinischen Labels. Verglichen wurden zwei Repräsentationsstrategien: (1) eine semantische Methode mit dimensionaler Reduktion und biomechanischen Merkmalen sowie (2) eine automatisierte Merkmalsextraktion. Die Modelle wurden auf den öffentlichen Daten trainiert und anschließend unverändert auf den eigenen klinischen Datensatz übertragen. Die Leistungsbewertung erfolgte mittels ungewichtetem F_1-Score.

Im Trainingssetting erreichten automatisierte Verfahren F_1-Werte von bis zu 0,68, während die semantische Methode 0,63 erzielte. In der externen klinischen Validierung hingegen zeigte sich eine hybride Strategie als am leistungsfähigsten ($F_1 = 0{,}53$), gefolgt von der semantischen ($F_1 = 0{,}48$) und der rein automatisierten Methode ($F_1 = 0{,}37$). Die Unterschiede waren statistisch signifikant ($p < 0{,}001$).

Die Ergebnisse belegen, dass klinisch fundierte Repräsentationen zu einer höheren Robustheit und Interpretierbarkeit führen. Damit liefern kontinuierlich aufgezeichnete, semantisch aufbereitete Sensorsignale eine tragfähige Grundlage für zukünftige Entscheidungsunterstützungssysteme zur individualisierten Therapieanpassung bei Parkinson. Für eine breite klinische Implementierung sind jedoch multizentrische Studien und standardisierte Protokolle erforderlich.

Die in dieser Arbeit entwickelten Methoden bilden bereits die Grundlage für zwei klinische Folgeprojekte, die auf vernetzten Sensorsystemen basieren und eine objektive, kontinuierliche Erfassung des Krankheitsverlaufs anstreben.

English Summary

This dissertation explores the methodological foundations, opportunities, and limitations of wrist-worn sensor systems for the automated detection of levodopa-induced dyskinesia in patients with Parkinson's disease. At its core, the study hypothesizes that semantically informed feature representations, aligned with clinical observation logic, offer better generalizability to unseen clinical data than purely data-driven approaches.

The analysis draws upon two publicly available datasets and a newly collected clinical dataset labeled by using standardized clinical criteria. Two representation strategies were compared: (1) a semantic method combining dimensionality reduction with biomechanical features, and (2) an automated time-series feature extraction approach. Models were trained on the public data and then applied unchanged to the clinical dataset. Performance was evaluated using the unweighted F_1-score.

In training, automated methods achieved F_1 scores up to 0.68, while the semantic approach reached 0.63. However, during external clinical validation, a hybrid strategy combining PCA with automated extraction performed best ($F_1 = 0.53$), followed by the semantic method ($F_1 = 0.48$) and the purely automated one ($F_1 = 0.37$). These differences were statistically significant ($p < 0.001$).

The results demonstrate that clinically grounded data representations yield higher robustness and interpretability. Consequently, continuous and semantically processed sensor signals provide a promising foundation for future decision support systems aimed at individualized therapy adjustment in Parkinson's care. However, broader clinical implementation will require larger multicenter studies and standardized data protocols.

The methods developed in this work already serve as the basis for two ongoing clinical follow-up projects that use networked sensor systems to enable objective, continuous monitoring of disease progression.

4 Literaturverzeichnis

Banos, O., Galvez, J.-M., Damas, M., Pomares, H., & Rojas, I. (2014). Window Size Impact in Human Activity Recognition. *Sensors (Basel, Switzerland)*, *14*(4), 6474–6499. https://doi.org/10.3390/s140406474

Bengio, Y., Courville, A., & Vincent, P. (2014). *Representation Learning: A Review and New Perspectives* (arXiv:1206.5538). arXiv. https://doi.org/10.48550/arXiv.1206.5538

Bloem, B. R., Post, E., & Hall, D. A. (2023). An Apple a Day to Keep the Parkinson's Disease Doctor Away? *Annals of Neurology*, *93*(4), 681–685. https://doi.org/10.1002/ana.26612

Christ, M., Braun, N., Neuffer, J., & Kempa-Liehr, A. W. (2018). Time Series FeatuRe Extraction on basis of Scalable Hypothesis tests (tsfresh – A Python package). *Neurocomputing*, *307*, 72–77. https://doi.org/10.1016/j.neucom.2018.03.067

Elm, J. J., Daeschler, M., Bataille, L., Schneider, R., Amara, A., Espay, A. J., Afek, M., Admati, C., Teklehaimanot, A., & Simuni, T. (2019). Feasibility and utility of a clinician dashboard from wearable and mobile application Parkinson's disease data. *Npj Digital Medicine*, *2*(1), Article 1. https://doi.org/10.1038/s41746-019-0169-y

Giannakopoulou, K.-M., Roussaki, I., & Demestichas, K. (2022). Internet of Things Technologies and Machine Learning Methods for Parkinson's Disease Diagnosis, Monitoring and Management: A Systematic Review. *Sensors*, *22*(5), Article 5. https://doi.org/10.3390/s22051799

Gundler, C., Zhu, Q. R., Trübe, L., Dadkhah, A., Gutowski, T., Rosch, M., Langebrake, C., Nürnberg, S., Baehr, M., & Ückert, F. (2023). A Unified Data Architecture for Assessing Motor Symptoms in Parkinson's Disease. *Studies in Health Technology and Informatics*, *307*, 22–30. https://doi.org/10.3233/SHTI230689

Hssayeni, M. D., Jimenez-Shahed, J., Burack, M. A., & Ghoraani, B. (2021). Dyskinesia estimation during activities of daily living using wearable motion sensors and deep recurrent networks. *Scientific Reports*, *11*(1), Article 1. https://doi.org/10.1038/s41598-021-86705-1

Jalloul, N. (2018). Wearable sensors for the monitoring of movement disorders. *Biomedical Journal*, *41*(4), 249. https://doi.org/10.1016/j.bj.2018.06.003

Kwon, D. K., Kwatra, M., Wang, J., & Ko, H. S. (2022). Levodopa-Induced Dyskinesia in Parkinson's Disease: Pathogenesis and Emerging Treatment Strategies. *Cells*, *11*(23), Article 23. https://doi.org/10.3390/cells11233736

Liashchynskyi, P., & Liashchynskyi, P. (2019). *Grid Search, Random Search, Genetic Algorithm: A Big Comparison for NAS* (arXiv:1912.06059; Version 1). arXiv. http://arxiv.org/abs/1912.06059

Pfister, F. M. J., Um, T. T., Pichler, D. C., Goschenhofer, J., Abedinpour, K., Lang, M., Endo, S., Ceballos-Baumann, A. O., Hirche, S., Bischl, B., Kulić, D., & Fietzek, U. M. (2020). High-Resolution Motor State Detection in Parkinson's Disease Using Convolutional Neural Networks. *Scientific Reports*, *10*(1), Article 1. https://doi.org/10.1038/s41598-020-61789-3

Pringsheim, T., Day, G. S., Smith, D. B., Rae-Grant, A., Licking, N., Armstrong, M. J., de Bie, R. M. A., Roze, E., Miyasaki, J. M., Hauser, R. A., Espay, A. J., Martello, J. P., Gurwell, J. A., Billinghurst, L., Sullivan, K., Fitts, M. S., Cothros, N., Hall, D. A., Rafferty, M., … Guideline Subcommittee of the AAN. (2021). Dopaminergic Therapy for Motor Symptoms in Early Parkinson Disease Practice Guideline Summary: A Report of the

AAN Guideline Subcommittee. *Neurology*, *97*(20), 942–957. https://doi.org/10.1212/WNL.0000000000012868

Pringsheim, T., Jette, N., Frolkis, A., & Steeves, T. D. L. (2014). The prevalence of Parkinson's disease: A systematic review and meta-analysis. *Movement Disorders: Official Journal of the Movement Disorder Society*, *29*(13), 1583–1590. https://doi.org/10.1002/mds.25945

Richter, D., Bartig, D., Muhlack, S., Hartelt, E., Scherbaum, R., Katsanos, A. H., Müller, T., Jost, W., Ebersbach, G., Gold, R., Krogias, C., & Tönges, L. (2019). Dynamics of Parkinson's Disease Multimodal Complex Treatment in Germany from 2010–2016: Patient Characteristics, Access to Treatment, and Formation of Regional Centers. *Cells*, *8*(2), Article 2. https://doi.org/10.3390/cells8020151

Sánchez-Fernández, L. P., Garza-Rodríguez, A., Sánchez-Pérez, L. A., & Martínez-Hernández, J. M. (2023). A Computer Method for Pronation-Supination Assessment in Parkinson's Disease Based on Latent Space Representations of Biomechanical Indicators. *Bioengineering*, *10*(5), Article 5. https://doi.org/10.3390/bioengineering10050588

Seo, S., Kim, Y., Han, H.-J., Son, W. C., Hong, Z.-Y., Sohn, I., Shim, J., & Hwang, C. (2021). Predicting Successes and Failures of Clinical Trials With Outer Product–Based Convolutional Neural Network. *Frontiers in Pharmacology*, *12*, 670670. https://doi.org/10.3389/fphar.2021.670670

Shawen, N., O'Brien, M. K., Venkatesan, S., Lonini, L., Simuni, T., Hamilton, J. L., Ghaffari, R., Rogers, J. A., & Jayaraman, A. (2020). Role of data measurement characteristics in the accurate detection of Parkinson's disease symptoms using wearable sensors. *Journal*

of NeuroEngineering and Rehabilitation, *17*(1), 52. https://doi.org/10.1186/s12984-020-00684-4

Motion Analysis Laboratory, D. O. P. M. & , Robert. (2019). MJFF Levodopa Response Study [Dataset]. Synapse. https://doi.org/10.7303/SYN20681023

Van Gerpen, J. A., Kumar, N., Bower, J. H., Weigand, S., & Ahlskog, J. E. (2006). Levodopa-associated dyskinesia risk among Parkinson disease patients in Olmsted County, Minnesota, 1976-1990. *Archives of Neurology*, *63*(2), 205–209. https://doi.org/10.1001/archneur.63.2.205

Wang, W. K., Chen, I., Hershkovich, L., Yang, J., Shetty, A., Singh, G., Jiang, Y., Kotla, A., Shang, J. Z., Yerrabelli, R., Roghanizad, A. R., Shandhi, M. M. H., & Dunn, J. (2022). A Systematic Review of Time Series Classification Techniques Used in Biomedical Applications. *Sensors*, *22*(20), Article 20. https://doi.org/10.3390/s22208016

5 Abkürzungsverzeichnis

UKE	Universitätsklinikum Hamburg-Eppendorf
PCA	Hauptkomponentenanalyse (Principal Component Analysis)
tsfresh	Time-Series Feature Extractors-Bibliothek (Python)
LID	Levodopa-induzierte Dyskinesie
PD	Morbus Parkinson (Parkinson's Disease)
MJFF	Michael J. Fox Stiftung
UPDRS	Unified Parkinson's Disease Rating Scale (Skala zur Bewertung der Parkinson-Krankheit)
LRS	Levodopa Response Study (Datensatz)
CIS-PD	Clinician Input Study (Datensatz)
PKB	Parkinson-Komplexbehandlung
PACMAN	Parkinson's Clinical Movement Assessment (Datensatz)
FHIR	Fast Healthcare Interoperability Resources (Standard für Gesundheitsdaten-Interoperabilität)
ML	Maschinelles Lernen
SMOTE	Synthetic Minority Over-Sampling Technique
AUPR	Fläche unter der Präzisions-Recall-Kurve (Area Under the Precision-Recall Curve)
DREAM	Digital Biomarker Evaluation and Analysis for Mobile Health Challenge (Studie)

6 Erklärung des Eigenanteils

Mein Anteil an der Publikationsdissertation ergibt sich gemäß der Autorenreihenfolge der Publikation. Er umfasste die Datenerhebung (PACMAN Datensatz), Literaturrecherche, Algorithmenentwicklung, Durchführung der Analyse und die Interpretation der Ergebnisse. Die vorliegende Publikation mitsamt ihrer Tabellen und Abbildungen wurde von mir eigenständig verfasst.

Das Studiendesign wurde von Prof. Dr. Frank Ückert, Prof. Dr. Monika Pötter-Nerger, PD Dr. Claudia Langebrake und Dr. Adrin Dadkhah entwickelt. Bei der Ausarbeitung der Fragestellung sowie der statistischen Datenanalyse wurde ich von Christopher Gundler, M.Sc. unterstützt. Die verwendete watchOS-Anwendung wurde von Sören Spiegel, M.Sc. programmiert. Die klinische Ausarbeitung der Studie sowie alle Auswertungsschritte wurden eigenständig von mir durchgeführt.

7 Eidesstattliche Versicherung

Ich versichere ausdrücklich, dass ich die Arbeit selbständig und ohne fremde Hilfe, insbesondere ohne entgeltliche Hilfe von Vermittlungs- und Beratungsdiensten, verfasst, andere als die von mir angegebenen Quellen und Hilfsmittel nicht benutzt und die aus den benutzten Werken wörtlich oder inhaltlich entnommenen Stellen einzeln nach Ausgabe (Auflage und Jahr des Erscheinens), Band und Seite des benutzten Werkes kenntlich gemacht habe. Das gilt insbesondere auch für alle Informationen aus Internetquellen.

Soweit beim Verfassen der Dissertation KI-basierte Tools („Chatbots“) verwendet wurden, versichere ich ausdrücklich, den daraus generierten Anteil deutlich kenntlich gemacht zu haben. Die „Stellungnahme des Präsidiums der Deutschen Forschungsgemeinschaft (DFG) zum Einfluss generativer Modelle für die Text- und Bilderstellung auf die Wissenschaften und das Förderhandeln der DFG“ aus September 2023 wurde dabei beachtet.

Ferner versichere ich, dass ich die Dissertation bisher nicht einem Fachvertreter an einer anderen Hochschule zur Überprüfung vorgelegt oder mich anderweitig um Zulassung zur Promotion beworben habe.

Ich erkläre mich damit einverstanden, dass meine Dissertation vom Dekanat der Medizinischen Fakultät mit einer gängigen Software zur Erkennung von Plagiaten überprüft werden kann.

Datum Unterschrift

8 Danksagung

Ich danke meinem Doktorvater, Prof. Dr. med. Frank Ückert, sehr herzlich für die Übernahme der Betreuung dieser Dissertation sowie für seine Unterstützung als Mentor, der mir jederzeit bei wissenschaftlichen Fragen zur Seite stand.

Mein besonderer Dank gilt Christopher Gundler, M.Sc., der mich als Leiter der Arbeitsgruppe für Angewandte KI im Gesundheitswesen (AAI) kontinuierlich bei wissenschaftlichen und fachlichen Fragestellungen unterstützt und mich während der gesamten Softwareentwicklung beratend begleitet hat.

Ebenso danke ich Prof. Dr. med. Monika Pötter-Nerger für die klinische Unterstützung im Rahmen der Studie sowie für die Bereitstellung der Patient:innenkohorte.

Mein Dank gilt außerdem Sören Spiegel, M.Sc., für die Entwicklung der watchOS-Anwendung zur Datenerhebung und Sara Tiedemann für die Unterstützung bei der Wissenschaftskommunikation.

Abschließend gilt mein tief empfundener Dank meiner Familie, insbesondere meiner Partnerin Christin Wienkamp, für ihren Rückhalt und ihre beständige Motivation.