Schlussbericht zum Verwendungsnachweis

Förderkennzeichen:	ZMI5-2522DAT11A
Titel des Vorhabens:	**Te**xtMining von **M**eldungstexten für **e**inheitliche **K**lassifikationen
Akronym:	**TeMeK**
Zuwendungsempfänger:	Krebsregister Baden-Württemberg (Konsortialführung ohne Zuwendung), Fraunhofer-Institut für Kommunikation, Informationsverarbeitung und Ergonomie (FKIE), Averbis GmbH, Albert-Ludwigs-Universität Freiburg / Universitätsklinikum Freiburg (UKF)
Projektleitung:	Prof. Dr. Marco Halber, Krebsregister Baden-Württemberg
Telefon:	0176 22224432
E-Mail:	halber@klr-krbw.de
Vorhabenart:	☒ Verbundvorhaben
Laufzeit:	01.09.2022 – 31.08.2025
Fördersumme:	1.123.410 €
Datum Berichterstellung:	27.02.2026

Gefördert durch:

Bundesministerium
für Gesundheit

Inhaltsverzeichnis

1. Zusammenfassung 3
2. Einleitung 4
   - 2.1 Ausgangslage / Hintergrund / Motivation des Projekts 4
   - 2.2 Forschungsfrage(n) 4
   - 2.3 Ziele des Projekts 4
   - 2.4 Projektstruktur 5
3. Erhebungs- und Auswertungsmethodik 6
4. Durchführung, Arbeits- und Zeitplan 8
5. Ergebnisse 10
   - 5.1 Arbeitspaket 1: Datenbereitstellung 10
   - 5.2 Arbeitspaket 2: Klassischer standardisierter Pathologie-Befund (AF 1) 10
     - 5.2.1 AP 2.1 Klassifikation der Standardentitäten (15 PM) 10
     - 5.2.2 AP 2.2 Konflikterkennung bei Standardentitäten (6 PM) 12
     - 5.2.3 AP 2.3 Organspezifische Zusatzmodule (15 PM) 12
     - 5.2.4 AP 2.4 Zusätzliche Klassifikationen (6 PM) 14
   - 5.3 Arbeitspaket 3: Erweiterter Molekularpathologischer Befund (AF 2) 15
     - 5.3.1 AP 3.1 – 3.4 Erkennung von Genen, Proteinmutationen, Ausprägung genetischer Varianten 15
     - 5.3.2 Arbeitspaket 3.5 Verbindung mit HAP2 24
   - 5.4 Arbeitspaket 4: Ergebnisvernetzung 25
6. Gender Mainstreaming Aspekte 26
   - 6.1 Organisatorische und strukturelle Gender Mainstreaming Aspekte 26
   - 6.2 Inhaltliche Gender Mainstreaming Aspekte 26
7. Partizipation 26
8. Diskussion der Ergebnisse, Gesamtbeurteilung 27
   - 8.1 Arbeitspaket 2 (AF1) 27
   - 8.2 Arbeitspaket 3 (AF2) 28
9. Verbreitung der Projektergebnisse, Öffentlichkeitsarbeit 31
10. Verwertung der Projektergebnisse (Fortführung, Nachhaltigkeit, Transferpotenzial) 32

1. Zusammenfassung

Problemhintergrund: Klinische Krebsregister sind auf valide und zeitnah verfügbare Daten angewiesen, sowohl für handlungsleitende Auswertungen zur Versorgungsverbesserung als auch für externe Qualitätssicherung. Ein erheblicher Teil der gemeldeten Informationen liegt jedoch als unstrukturierter Freitext vor, insbesondere in Pathologie- und molekularpathologischen Befunden. Die manuelle Extraktion, Validierung und Kodierung dieser Informationen ist personalintensiv, fehleranfällig und angesichts steigender Datenmengen langfristig nicht skalierbar. Insbesondere Inkonsistenzen zwischen Freitext und strukturierten Angaben sowie das dynamisch wachsende und hochspezialisierte Vokabular der Molekularpathologie stellen die Register vor erhebliche Herausforderungen. Mit der Weiterentwicklung des Datensatzes der Arbeitsgemeinschaft Deutscher Tumorzentren (ADT) und der Gesellschaft der epidemiologischen Krebsregister in Deutschland (GEKID) und der Ausweitung der Meldepflicht molekularer Befunde hat die Bedeutung dieser Daten stark zugenommen. Molekulare Marker sind heute zunehmend zentral für Klassifikation, Prognose und Therapieentscheidung und bilden die Grundlage moderner, personalisierter Onkologie. Gleichzeitig fehlen in den Krebsregistern manchmal die personellen und fachlichen Ressourcen, um diese Informationen vollständig und konsistent aus Freitexten zu erschließen; ohne geeignete technische Unterstützung drohen relevante Informationen ungenutzt zu bleiben oder verspätet verfügbar zu sein.

Zentrale Forschungsfragen und Ziele: Untersucht wurde, welche Methoden der Künstlichen Intelligenz (KI) geeignet sind, solch komplexe medizinische Freitexte valide, reproduzierbar und effizient zu analysieren. Ziel ist es, befundrelevante Informationen automatisiert zu extrahieren, Inkonsistenzen zu erkennen, und eine teil- bis vollautomatisierte Verarbeitung zu ermöglichen. Übergeordnetes Ziel dabei ist die bundesweit einheitliche Bereitstellung qualitativ hochwertiger Krebsregisterdaten für Auswertung, Forschung und Qualitätssicherung.

Herangehensweise: Das Projekt verfolgt zwei eng verzahnte Anwendungsfelder. Im ersten werden klassische pathologische Befunde analysiert, indem Deep-Learning-basierte Sprachmodelle mit Convolutional Neural Networks (CNN; faltende neuronale Netze) kombiniert und durch regelbasierte Verfahren ergänzt. Dadurch sollen Standardentitäten des deutschen onkologischen Basisdatensatzes sowie organspezifische Zusatzmodule zuverlässig erkannt, validiert und dokumentiert werden. Das zweite Anwendungsfeld fokussiert auf molekularpathologische Befunde. Hier werden zeichen-, regel- und kontextbasierte Verfahren eingesetzt, um Gene, Varianten und proteinbezogene Veränderungen zu identifizieren, zu klassifizieren und in standardisierte Notationen zu überführen. Beide Anwendungsfelder werden in einer modularen Microservice-Architektur umgesetzt, die eine flexible Integration in unterschiedliche Krebsregister erlaubt und schrittweise erweitert werden kann.

Wichtigste Ergebnisse: Die grundsätzliche Machbarkeit und der praktische Nutzen der Methodik konnten belegt werden. Im Bereich der klassischen Pathologiebefunde konnten die Erkennungsraten zentraler Entitäten, insbesondere der Diagnose-, Morphologie- und Tumor-Node-Metastasis-Kodierung (TNM-Kodierung), signifikant verbessert werden. Für ausgewählte Tumorentitäten, etwa die Prostata, wurden organspezifische Zusatzmerkmale soweit automatisiert extrahiert, dass eine weitergehende Kodierunterstützung erreicht wurde. Im Bereich der Molekularpathologie wurde ein stabiler Workflow zur Identifikation, Kontextualisierung und Aggregation molekularer Marker etabliert. Relevante genetische Aussagen können aus umfangreichen Befunden extrahiert, konsolidiert und nachvollziehbar für die Dokumentation aufbereitet werden. Die Zwischenergebnisse zeigen, dass auch hochkomplexe molekulare Informationen mit vertretbarem Aufwand strukturiert erschlossen werden können und damit

erstmals systematisch für Krebsregisterauswertungen verfügbar werden. TeMeK schafft damit eine belastbare methodische und technische Grundlage für eine nachhaltige Verbesserung der Krebsregisterdaten. Die entwickelten Verfahren eröffnen Perspektiven für eine weitergehende Automatisierung, für registerübergreifende Harmonisierung sowie für den Einsatz in angrenzenden Bereichen der onkologischen Versorgungsforschung und Qualitätssicherung.

2. Einleitung

2.1 Ausgangslage / Hintergrund / Motivation des Projekts

Die Daten der klinischen Krebsregister sind die Grundlage handlungslenkender Auswertungen, die zur Verbesserung der Versorgung an Krebs erkrankter Personen beitragen. Außerdem bilden die Daten die Grundlage für externe Qualitätssicherung für die Leistungserbringerinnen und Leistungserbringer der onkologischen Versorgung. Daraus leiten sich höchste Anforderungen an die Datenqualität und die zeitnahe Verfügbarkeit der Daten ab. Die Meldungen an die klinischen Krebsregister enthalten in großem Umfang unstrukturierten Freitext. Die Extraktion der relevanten Daten inklusive komplexer Plausibilitätsprüfungen ist mit einem sehr hohen manuellen Aufwand verbunden.

2.2 Forschungsfrage(n)

Die Fragestellung des Projekts war darum: Welche Methoden der Künstlichen Intelligenz (KI) können eingesetzt und wie können sie angepasst und weiterentwickelt werden, um eine einheitliche, valide und effiziente Informationsextraktion aus komplexem Freitext zu ermöglichen? Hierzu wurde im Projekt auf die folgenden zwei Anwendungsfelder AF1 und AF2 fokussiert:

Anwendungsfeld 1: Pathologie-Befunde: Der freitextliche Pathologiebefund stellt die robuste inhaltliche Referenz ("Goldstandard") für die valide Beschreibung des Tumors und seiner Ausbreitung dar. Daraus abgeleitete Verschlüsselungen in Form von Zeichencodes sind fehleranfällig und bedürfen der Validierung. Mit vielen hunderttausenden Tumormeldungen jährlich ist die rein manuelle Validierung und Tumordokumentation auf Basis von freitextlichen Pathologiebefunden nicht zu bewältigen, so dass es zu Dokumentationsengpässen und einem Befundstau in den Klinischen Krebsregistern kommt. Daher sollten in AF1 Verfahren aus dem Bereich des Text Mining und Natural Language Processing (NLP; Verarbeitung natürlicher Sprache) erweitert und bereitgestellt werden, mit denen die Tumordokumentation unterstützt, Entlastung bei der manuellen Dokumentation geschaffen und der Automatisierungsgrad bei der Erfassung erhöht werden kann.

Anwendungsfeld 2: Molekularpathologischer Befund: Der molekularpathologische Befund präzisiert die Beschreibung der Tumorbiologie. Er spielt eine zunehmend herausragende Rolle bei der Risikostratifizierung und der Indikationsstellung einer zielgerichteten, individualisierten Therapie. Der freitextliche molekularpathologische Befund enthält ein dynamisch wachsendes und komplexes Lexikon, dessen Interpretation häufig Fachwissen erfordert. Daher sollten in AF2 Verfahren aus dem Bereich des Text-Mining/Natural Language Processing entwickelt und bereitgestellt werden, die die Erschließung der relevanten Inhalte aus dem Freitext erleichtern und eine Interpretation der Inhalte auf fachlich qualifiziertem Niveau unterstützen.

2.3 Ziele des Projekts

Das übergeordnete allgemeine Ziel, welches über die Projektlaufzeit hinaus bestehen bleibt, ist die schnelle Bereitstellung von qualitativ hochwertigen Daten, die bundesweit einheitlich für eine Zusammenführung am Zentrum für Krebsregisterdaten und für eine belastbare externe Qualitätssicherung taugen. Das konkrete Ziel für die Projektlaufzeit war es am Beispiel der

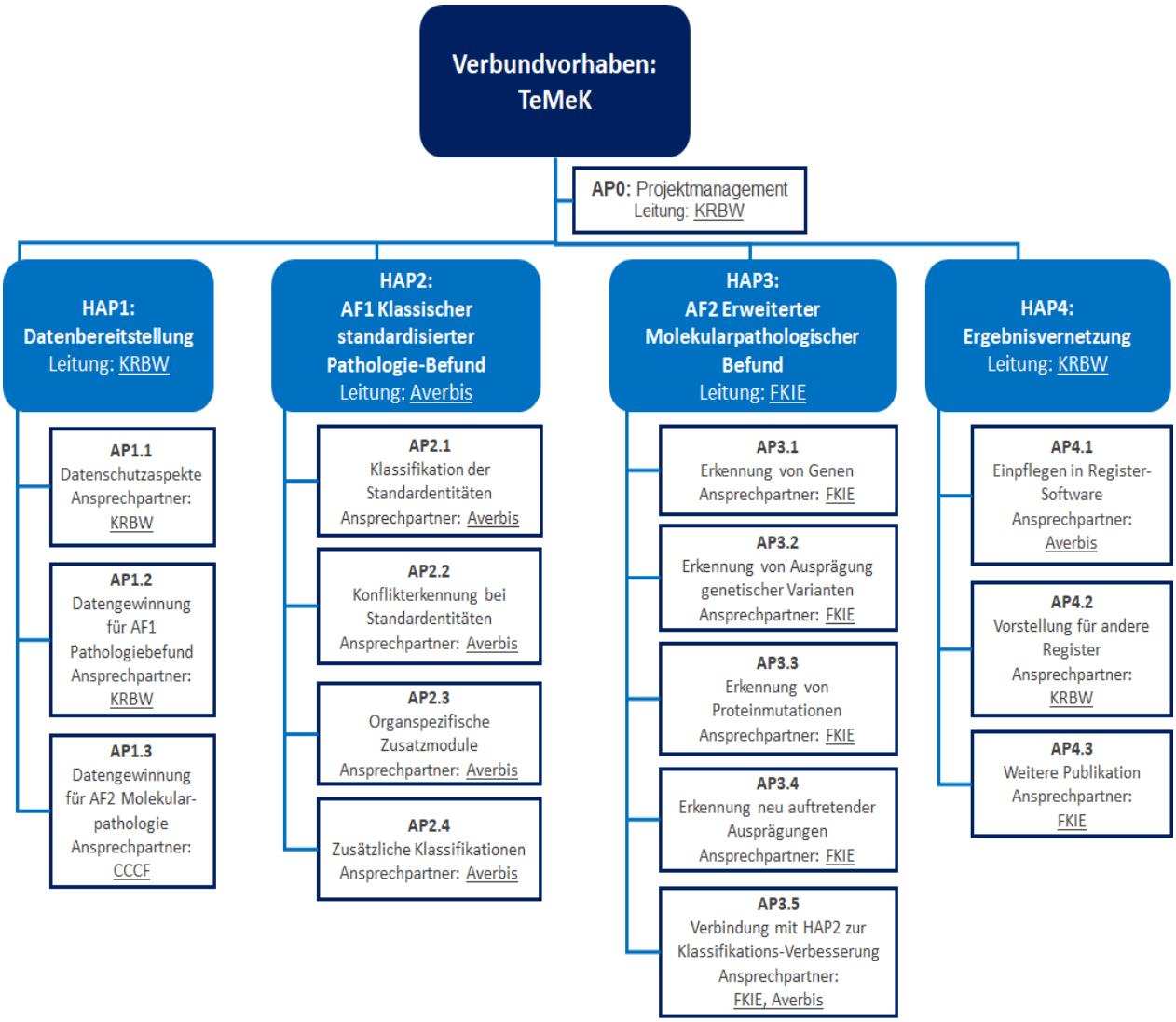
hochrelevanten Anwendungsfelder AF1 und AF2 zu erforschen und demonstrieren, wie Verfahren aus dem KI-gestützten Text Mining/Natural Language Processing zur Erreichung es allgemein Zieles beitragen können:

Anwendungsfeld 1: Pathologie-Befunde: Projektziel für diesen Anwendungsfall war die möglichst vollständige Extraktion von befundrelevanten Merkmalen aus Pathologieberichten. Dabei sollten die im einheitlichen onkologischen Basisdatensatz gemäß § 65c des Fünften Buches Sozialgesetzbuch (SGB V) der ADT/GEKID genannten Felder sowie die Felder der organspezifischen Ergänzungsmodule zuverlässig erkannt werden. Hier galt es Verfahren zu entwickeln, die die Inkonsistenzen im Freitext und die Inkonsistenzen zwischen dem Freitext und Einträgen in strukturierten Feldern detektieren, die Validierung unterstützen und die effektive Dokumentation der validierten Inhalte bis hin zu einer automatisierten Verarbeitung ermöglichen. Darüber hinaus sollte eine Grundlage geschaffen werden, um befundrelevante Zusatzinformationen zur weiteren Verarbeitung in den Krebsregistern zu extrahieren. Das über die Projektlaufzeit hinausgehende Ziel bleibt, die Krebsregister durch diese Arbeiten mittelfristig auf vielfältige Weise zu entlasten: a) durch Dokumentationsunterstützung für Kodierfachkräfte, b) Validierung durch Abgleich mit weiteren strukturierten Befunden mit Änderungsvorschlägen, und c) Effizienzsteigerung durch Optimierung des Prüfprozesses pathologischer Befundangaben bei Routine-Befund-Klassifikation.

Anwendungsfeld 2: Molekularpathologischer Befund: Projektziel für diesen Anwendungsfall die Erschließung bisher nicht zugänglicher molekularpathologischer Informationen in hoher Qualität und hohem Strukturierungsgrad, um sie für abgeleitete Auswertungen verfügbar zu machen. Dabei sollten genetische Informationen, Hauptmutationen und Modifikationen erkannt und als einzelne Informationseinheiten zugreifbar gemacht werden: a) Häufige Mutationen sollten in einer Form extrahiert und normalisiert werden, die es erlaubt, dass sie in entsprechende strukturierte Felder befüllt werden können; b) seltenere Mutationen sollten in nicht-strukturierten Feldern erfasst werden (ADT/GEKID-Spezifikation 3.0.0, Feld „Genetik“). Das über die Projektlaufzeit hinausgehende Ziel bleibt, für die Krebsregister diagnoserelevantes Wissen extrahieren und bereitstellen zu können, da insbesondere in der neuen Revision der Internationalen Klassifikation der Krankheiten für die Onkologie (ICD-O 3.2) molekular-pathologische Informationen eine wachsende Rolle zur Kodierung der Morphologie spielen.

2.4 Projektstruktur

Das Projekt wurde mittels fünf Hauptarbeitspaketen (HAP) strukturiert, wobei die oben genannten zwei Anwendungsfelder AF1 und AF2 in den zwei unabhängigen HAP 2 und 3 bearbeitet und mit dem HAP 1 zur Datenbereitstellung verzahnt wurden.



3. Erhebungs- und Auswertungsmethodik

Bzgl. der Erhebungs- und Auswertemethodik wurde für die beiden Anwendungsfelder AF1 und AF2 in den einzelnen Teilarbeitspaketen (AP) grundsätzlich analog vorgegangen. Zur Bewertung der Zielerreichung wurden verschiedene methodische Ansätze eingesetzt.

Arbeitspaket 2 (AF1): Im Mittelpunkt stand die quantitative Messung der Vollständigkeit und Qualität der automatisierten Extraktion befundrelevanter Merkmale sowie die Überprüfung ihrer Konsistenz mit vorhandenen Kodierungen.

Die Vollständigkeit der extrahierten Merkmale wurde durch einen strukturierten Abgleich mit den relevanten Feldern des onkologischen Basisdatensatzes (oBDS) geprüft. Hierbei wurde analysiert, inwieweit alle erforderlichen Informationen aus den unstrukturierten Freitextbefunden identifiziert und in strukturierter Form abgebildet werden können.

Zur Beurteilung der Extraktionsqualität wurden die Metriken Recall, Precision und der daraus abgeleitete F1-Score berechnet. Diese Kennzahlen wurden sowohl für einzelne diagnostische Merkmale (z. B. den Diagnosecode) als auch aggregiert auf Ebene spezifischer Tumorentitäten ausgewertet. Dadurch konnte die Leistung der Extraktionsverfahren differenziert bewertet und Optimierungspotenziale identifiziert werden.

Ein weiterer Auswertungsschwerpunkt lag auf der Detektion von Inkonsistenzen in den vorhandenen Kodierungen der Daten. Hierzu kamen deskriptive statistische Verfahren, Kreuztabellenanalysen sowie Recall-, Precision- und F1-Auswertungen zum Einsatz. Diese Analysen wurden teils auch differenziert nach einzelnen Krebsregistern durchgeführt, um regionale Unterschiede und potenzielle systematische Abweichungen sichtbar zu machen.

Schließlich wurde geprüft, inwieweit die entwickelten Verfahren eine automatisierte Verarbeitung ermöglichen. Dies wurde insbesondere durch die erreichte Vollständigkeit und Qualität der Extraktion bewertet, die entscheidende Voraussetzungen für eine belastbare und skalierbare Nutzung in der Routineverarbeitung darstellen.

Arbeitspakete 3.1 – 3.4 (AF2): Neben den bereits für AF1 beschriebenen quantitativen Analysen fanden in AF2 für die Unterarbeitspakete AP 3.1 – AP 3.4 zusätzlich qualitative Analysen statt, d.h. es wurde für die unterschiedlichen möglichen Fehlerarten untersucht, welche Fehlerursachen in der Praxis auftreten. Diese Erkenntnis ist wichtig für eine mögliche Weiterentwicklung der Textmining-Verfahren, denn einige Fehler können potenziell durch algorithmische Verfeinerungen der Textmining-Verfahren ausgeräumt werden, wohingegen andere Fehler sich nur durch eine Harmonisierung von Notation und Sprache in molekularpathologischen Befunden vermeiden lassen.

Datengrundlage für die quantitative und qualitative Auswertung der AP 3.1 – AP 3.4 war ein Korpus von 21.000 anonymisierten Meldungen aus dem Zeitraum 2020 – 2023 von 53 unterschiedlichen meldenden Einrichtungen und mit Tumoren aus ca. 25 unterschiedlichen Lokalisationen. Der Umfang des Testkorpus ist wie folgt:

- ca. 21.000 Meldungsdokumente (mit teils konkatenierten Befunden)
- ca. 38.000 einzelne Befunde (nach Auflösung der Konkatenation)
- ca. 700.000 Nennungen von Biomarkern in den Befunden, wobei ca. 90 unterschiedliche Biomarker betrachtet wurden.

Der Korpus wurde im Hinblick auf die Korrektheit bzw. Häufigkeit von Fehlklassifikationen (False-Positives, False-Negatives) der Textmining-Verfahren für unterschiedliche Aufgaben untersucht. Die Aufgaben bildeten hierbei jeweils ein Teilziel und waren wie folgt definiert:

- **Teilziel AF2.1:** Erkennung von Biomarkern (AP 3.1), so dass auch neu zu dokumentierende Biomarker jederzeit bearbeitet werden können (AP 3.4)
- **Teilziel AF2.2:** Extraktion unterschiedlicher spezifischer Patterns, u.a. Varianten der komplementären Desoxyribonukleinsäure (cDNA), Proteinvarianten sowie Exone (AP 3.3), so dass auch neu auftretende Varianten detektiert werden können (AP 3.4)
- **Teilziel AF2.3:** Erkennung von zu dokumentierenden Aussagen zu Genen (Zwischenziel von AP 3.2 und 3.4)
- **Teilziel AF2.4:** Klassifikation der Ausprägung möglicher Gen-Varianten (AP 3.2) unter Verwendung von generischen Ansätzen, so dass auch neu auftretende Ausprägungen damit erkannt werden können (AP 3.4)

Arbeitspaket 3.5 (AF2): Zur Validierung der im Rahmen von AF2 AP3.5 entwickelten Extraktionsverfahren wurden nach der Bereitstellung bzw. Erstellung eines Goldstandards dieselben Metriken (Recall, Precision und F1-Score) wie in AF1 verwendet. Sie ermöglichten eine differenzierte Bewertung der Extraktionsqualität für die identifizierten molekularpathologischen Merkmale, insbesondere spezifischer Biomarker und trugen damit zur Optimierung der Verfahren bei.

4. Durchführung, Arbeits- und Zeitplan

Arbeitspaket 1: Für das Arbeitspaket 1 waren seitens Averbis zwar keine eigenen Arbeitsaufwände vorgesehen, jedoch erwies sich eine enge Abstimmung mit den Krebsregistern, insbesondere im Hinblick auf die Anforderungen aus HAP 2, als notwendig.

Die Beiträge des Comprehensive Cancer Center Freiburg (CCCF) erfolgten im Rahmen der geplanten Aufwände. Die intensive Abstimmung mit den Projektpartnern erfolgte insbesondere in der Phase der Identifizierung und Priorisierung der Datenelemente, beispielsweise anhand der Pathways der verschiedenen Entitäten.

Für das **Arbeitspaket 1.1**, die Sicherstellung der Datenschutzaspekte, wurde seitens der Klinischen Landesregisterstelle Baden-Württemberg GmbH (KLR) die Prüfung der zur Verfügung gestellten Trainingsdaten vorgenommen. Ziel der Maßnahmen war der Ausschluss der Weitergabe personenidentifizierender Angaben an die Analysepartner bzw. die weiteren Projektpartner. In den vertraglichen Regelungen zwischen den Projektpartnern zur Datenweitergabe und Nutzung wurden weitere Sicherheitsmaßnahmen verabredet, um die Risiken von Reidentifizierung zu unterbinden. Der Datenfluss war zwischen jeweiligem Krebsregister und jeweiligem Praxispartner (Averbis, FKIE) getrennt, so dass die Sichtbarkeit von Trainingsdaten nur im zugehörigen Register und beim bestimmten Praxispartner gegeben war und kein weiterer Zugriff erfolgen konnte. Die Datenübermittlung erfolgte gewohnt verschlüsselt (256bit). Weiter wurde im Rahmen der Zielsetzung von Datensparsamkeit auf Freitextbefunde mit einem Mindestinhalt geachtet, damit die zulässigen rein strukturierten Meldungen nicht in Datenlieferungen eingehen konnten. Dieses Vorgehen ist bei vorbestehenden Datenlieferungen in anderen Projekten etabliert worden, so dass hierfür keine zeitlichen Verzögerungen verursacht wurden.

Für das **Arbeitspaket 1.2** Datengewinnung wurde auf vorliegende Pathologiefreitexte zurückgegriffen, die in den beteiligten Krebsregistern zur Verarbeitung vorlagen. Es erfolgten abgesprochene Selektionsalgorithmen, nach denen einheitlich über die beteiligten Krebsregister Datensatzausschnitte (Pathologiefreitexte) in ausreichender Anzahl zur Verfügung gestellt wurden. Für das Teilprojekt der Weiterentwicklung mit AVERBIS: Parallel zur Textlieferung wurden die strukturierten (medizinischen) Soll-Vorgaben als Ziel- und Lernwerte in eine über eindeutige Pseudonyme mit den Textdateien zusammenführbaren Tabelle übergeben. Da es sich in den Werten und Pseudonymen nur um med. Klassifikationen bzw. speziell vergebene IDs handelt, ist dadurch der Datenschutz gewährleistet. Im Teilprojekt der Bereitstellung an das FKIE wurden eigens dafür markierte Pathologiefreitexte anonymisiert bereitgestellt, welche genetische Marker enthielten. Eine Datei mit Ergebniswerten wurde nicht übergeben. Der Abgleich der Ergebnisse dazu fand in der KLR bzw. in den beteiligten Krebsregistern zu den jeweiligen Daten statt. Bei der Bereitstellung der Daten für die Analyse in AP 1.2 kam es krankheits- bzw. urlaubsbedingt zu geringfügigen Verzögerungen, die jedoch den Gesamtprojektfortschritt nicht maßgeblich beeinflussten.

Für das **Arbeitspaket 1.3** (CCCF) erfolgte zum Abgleich der in den Landeskrebsregistern dokumentierten Informationen mit den im Tumordokumentationssystem kodierten Daten ein Austausch zu den laufenden Projekten der Universitätsklinika im Bereich Molekulargenetik, unter anderem zu genomDE sowie zum Deutschen Netzwerk für Personalisierte Medizin (DNPM; Zentren für Personalisierte Medizin: ZPM). Auf Grundlage dieser Arbeitsergebnisse wurde im Rahmen der Gruppenarbeit ein umfangreiches Dokument als Wissensspeicher zum Thema Molekularpathologie erstellt und abgestimmt. Zudem wurde die Abbildung molekulargenetischer Informationen in den Landeskrebsregistern diskutiert. Darüber hinaus wurden Regelungen zur

Dokumentation molekulargenetischer Informationen gemäß dem onkologischen Basisdatensatz erarbeitet.

Arbeitspaket 2 (AF1): Im Verlauf von AP 2 waren keine Anpassungen des Zeitplans und der Ziele erforderlich. Obwohl die Bereitstellung der Daten nicht durchgehend im ursprünglich vorgesehenen Zeitrahmen realisieren ließ, konnte durch vorausschauende Planung bereits parallel an den verschiedenen Teilarbeitspaketen gearbeitet werden. Besonders bei den organspezifischen Zusatzmodulen gelang es, die Extraktion neuer Merkmale exemplarisch anhand der Dokumente eines Krebsregisters zu implementieren. Mit Bereitstellung weiterer Daten aus anderen Registern, konnten diese zur Validierung und Generalisierung der Komponenten herangezogen werden. Die Meilensteine wurden im Allgemeinen fristgerecht erreicht. Eine Ausnahme bildet jedoch Meilenstein M2.3 *Zusatzmodule klassifiziert*: Hier wurde eines der vier geplanten organspezifischen Zusatzmodule nicht abgeschlossen.

Arbeitspaket 3 (AF2): Im Verlauf von AP 3 waren ebenfalls keine Anpassungen des Zeitplans oder der Ziele erforderlich. Zwar war bereits zu Projektbeginn klar, dass die Bereitstellung von umfangreichen und detailliert annotierten Trainingsdaten aufgrund der Neuheit der Dokumentationsvorgaben zu genetischen Informationen erst spät im Projekt möglich werden würde, jedoch konnte eine negative Auswirkung auf den Projektverlauf erfolgreich vermieden werden: Direkt ab Projektbeginn fand ein kontinuierlicher Austausch und Wissenstransfer zwischen den Fachleuten der Krebsregister und den Entwicklungsteams der Text-Mining-Verfahren beim FKIE statt, so dass diese befähigt wurden, Trainingsdaten in der jeweils benötigten Spezifizierung, welche sich zum Teil auch erst dynamisch und kurzfristig in Abhängigkeit einer gerade verfolgten Extraktionsaufgabe ergab, selbst zu erzeugen. Der durchgehend engen Rückkopplung mit AP 1 ist es zu verdanken, dass die Meilensteine eingehalten, die Teilarbeitspakete wie geplant durchgeführt und die Ziele erreicht werden konnten. Die Unterarbeitspakete wurden insgesamt sequenziell, jeweils mit kleinen Überlappungen abgearbeitet, und hierbei jeweils auf das Teilergebnis des vorangegangenen Unterarbeitspakets zurückgegriffen. Zunächst wurde ein Ansatz für die systematische Erkennung von Genen umgesetzt (AP 3.1), darauf basierend ein Korpus für die Erkennung der Ausprägung genetischer Varianten (AP 3.2) und der Erkennung von Proteinmutationen (AP 3.3) dynamisch aufgebaut und dann für die Entwicklung entsprechender Klassifikatoren verwendet. Die Erkenntnisse der AP 3.1 - 3.3 wurden im Anschluss generalisiert, sodass auch neu auftretende Ausprägungen (AP3.4) i.d.R. ohne weitere Anpassung der Verfahren oder lediglich durch eine aufwandsarme Ergänzung von Wortschätzen erkannt werden können. Die Arbeiten von Averbis in HAP 3 (AP3.5) konzentrierten sich auf die zweite Projekthälfte. Dennoch fanden bereits in der ersten Hälfte vorbereitende Abstimmungen zwischen den Projektpartnern sowie erste Vorarbeiten statt.

AP 4: Es zeigte sich, dass das Einpflegen in die Registersoftware (AP 4.1) nicht als einmalige Maßnahme, sondern wiederholt, mindestens einmal pro Projektjahr, erforderlich war. Gleiches gilt für die Vorstellung der Entwicklungen bei weiteren Registern (AP 4.2), die ebenfalls mehrfach stattfand, um jeweils den aktuellen Projektstand sowie neue Ergebnisse zu präsentieren.

5. Ergebnisse

5.1 Arbeitspaket 1: Datenbereitstellung

Im Rahmen des AP1 sollten nach Klärung von Datenschutzaspekten die Datengewinnung in den beteiligten Krebsregistern und im CCCF abgestimmt und ermöglicht werden.

Aufgrund der Vorerfahrungen der KLR zur Datenbereitstellung für Analysezwecke konnten bekannte Prozesse und Maßnahmen übernommen und entsprechend der Zielsetzung umgesetzt werden, um eine datensparsame und datenschutzkonforme Bereitstellung an AVERBIS bzw. FKIE vorzunehmen. Neben der Datenbereitstellung wurde auch der Übertragungsweg per verschlüsselter und passwortgeschützter Bereitstellung und Passwortübergabe über einen davon unabhängigen Übertragungsweg sichergestellt. Data Sharing Agreements regelten die rechtlichen Pflichten der teilnehmenden Einrichtungen.

Die Abstimmung zu den Selektionskriterien für die bereit zu stellenden Daten (anonymisierte Freitextberichte der Pathologien) fand jeweils multilateral zwischen den Krebsregistern und Forschungspartnern AVERBIS bzw. FKIE statt. Die anschließenden Datenbereitstellungen erfolgten in einem Zeitfenster von 3-6 Wochen nach Festlegung der exakten Selektionskriterien. Über die Absprachen wurde auch sichergestellt, dass zur Zielsetzung entsprechend Daten bzw. Ergebnisdaten als Lerndaten verfügbar und in Gesamtsumme zur Verfügung gestellt werden konnten. Im Projektzeitraum wurden seitens der KLR und der beteiligten Krebsregister 5 Zyklen an Datenlieferungen an das FKIE sowie 6 Zyklen an Datenlieferungen an AVERBIS umgesetzt. Durch die Bereitstellung von Daten aus unterschiedlichen Krebsregistern wurde der Vielzahl individueller Pathologiefreitextformulierungen Rechnung getragen, welches die Einsatzmöglichkeit der entwickelten Modelle verbreitert und gewährleistet.

Das CCCF war intensiv in die Selektion und Priorisierung der relvanten Marker involviert. Nachfolgend wurde das erarbeitete Wissen intern genutzt, um die Bereitstellung und Dokumentation der relevanten Datenelemente sicherzustellen. Basierend darauf wurde die Schnittstellenimplementierung zwischen Tumordokumentationssystem Onkostar und Health Discovery vorangetrieben. Zeitweise technische Schwierigkeiten bei Aktualisierung der Software im klinischen Netzwerk konnten schlussendlich überwunden werden. Die in der Testung funktionale Lösung bietet Raum für weitere Optimierungen und ist mit zusätzlichem Entwicklungsaufwand auf andere Kliniken übertragbar.

5.2 Arbeitspaket 2: Klassischer standardisierter Pathologie-Befund (AF 1)

5.2.1 AP 2.1 Klassifikation der Standardentitäten (15 Personenmonate [PM])

Im Rahmen dieses Arbeitspakets sollten dokumentationsrelevante Tumorinformationen gemäß des onkologischen Basisdatensatzes (oBDS) automatisiert aus freitextlichen Pathologiebefunden extrahiert werden. Dabei standen insbesondere die folgenden Merkmale als Standardentitäten im Fokus:

- Diagnose
- Topographie
- Morphologie
- Seitenlokalisation
- Grading
- Residualstatus
- TNM-Klassifikation inkl. Symbolen und Kategorien
- Anzahl der befallenen & entnommenen (Sentinel-)Lymphknoten

Auf Basis bestehender Vorarbeiten von Averbis war die grundsätzliche Extraktion dieser Merkmale bereits vor Projektbeginn technisch möglich. Im Projektverlauf wurden die Extraktionskomponenten jedoch sowohl technologisch überarbeitet als auch qualitativ deutlich verbessert. Durch die umfassenden Datenbereitstellungen im Arbeitspaket 1.2 konnten die Modelle zudem auf eine deutlich breitere Grundlage gestellt werden: War eine zuverlässige Extraktion zu Projektbeginn nur für wenige Tumorentitäten möglich, so gelingt sie nun für nahezu alle Entitäten in überzeugender Qualität.

Ein direkter Vergleich der F1-Scores zu Projektbeginn und -ende ist aufgrund der geänderten Datengrundlage nur eingeschränkt aussagekräftig. Dennoch sind F1-Scores gute Indikatoren, um klare Fortschritte in Teilschritten der Entwicklung zu dokumentieren. So konnte der F1-Score für den Diagnosecode im Datensatz aus dem ersten Projektjahr um 0,038 Punkte gesteigert werden, im finalen Datensatz kam eine weitere Verbesserung um 0,017 hinzu. Der aktuelle F1-Score für dieses zentrale Merkmal liegt somit bei 0,849, wobei hier der Code der Internationalen statistischen Klassifikation der Krankheiten und verwandter Gesundheitsprobleme, 10. Revision (ICD-10) bis zur letzten Stelle berücksichtigt wird (z. B. *C50.9*). Bei einer auf die Diagnosegruppe (z. B. *C50*) beschränkten Betrachtung, ergibt sich ein nochmals höherer F1-Score von 0,948.

Da die Verteilung der Befunde nach Tumorentität im Testdatensatz stark variiert, schwanken auch die zugehörigen F1-Scores. Die folgende Tabelle zeigt exemplarisch die Ergebnisse für die häufigsten Tumorentitäten:

Tumorentität	**Anz. Dokumente im Testset**	**Proz. Anteil Testset (n = 42 747)**	**F1-Score Diagnose-code**	**F1-Score Diagnose-gruppe**
Mamma	5.754	13,46%	0,838	0,989
Prostata	5.648	13,21%	0,998	0,998
Malignes Melanom	4.252	9,95%	0,884	0,987
Kolorektales Karzinom	3.454	8,08%	0,832	0,948
Nicht-melanotischer Hautkrebs	3.163	7,40%	0,880	0,969
Lunge	2.669	6,24%	0,833	0,973

Bezieht man ausschließlich diese sechs häufigsten Tumorentitäten ein (etwa 58 % des gesamten Testdatensatzes), ergeben sich ein F1-Score von 0,886 für die Diagnosecodes sowie 0,981 für die Diagnosegruppen.

Neben der Diagnose konnten auch bei allen weiteren Standardentitäten Qualitätssteigerungen erzielt werden. Besonders hervorzuheben ist die TNM-Klassifikation. Die F1-Scores für die Kategorien *Tumor*, *Node* und *Metastasis* konnten - insbesondere durch den Einsatz von Machine Learning (ML; maschinelles Lernen) - jeweils um ca. 0,25 Punkte gesteigert werden. Die aktuellen Werte liegen bei:

- Tumor (T): 0,934
- Node (N): 0,962
- Metastasis (M): 0,963

Im Rahmen dieser Weiterentwicklungen wurden zudem neue technologische Ansätze erprobt. Bereits im ersten Projektjahr erfolgte der Wechsel von klassischen Support-Vector-Machine-

Modellen (SVM; Stützvektormaschinen) hin zu Convolutional Neural Networks (CNNs) für fünf Standardentitäten. Aufgrund der hohen Relevanz einzelner Merkmale (insbesondere der Diagnose) sowie inhaltlicher Abhängigkeiten zwischen bestimmten Entitäten wurde zusätzlich untersucht, ob Multitask-CNNs eine höhere Vorhersagegüte liefern könnten. Hierzu wurden kombinierte Modelle für verschiedene Feature-Sets trainiert. Die Ergebnisse zeigten allerdings keine signifikanten Verbesserungen gegenüber den Einzelmodellen. Zudem führte das häufige Fehlen einzelner Merkmalsannotationen zu einem relevanten Datenverlust, sodass von einem Einsatz im Routinebetrieb vorerst abgesehen wurde. Bei weiterer Optimierung und Vervollständigung der Datengrundlage bieten die Multitask-CNNs aber dennoch vielversprechende Möglichkeiten für die Zukunft.

5.2.2 AP 2.2 Konflikterkennung bei Standardentitäten (6 PM)

Ziel dieses Teilarbeitspakets war es, die Extraktionsergebnisse systematisch zu analysieren, um Unterschiede in der Kodierung zwischen sowie innerhalb der beteiligten Krebsregister zu identifizieren. Diese Analyse wurde durchgeführt und diente als zentrale Grundlage für gezielte Verbesserungen in anderen Arbeitspaketen - insbesondere bei der Anpassung und dem Training der Extraktionsmodelle.

Im Verlauf der Analyse zeigten sich je nach Merkmal zum Teil deutliche Unterschiede in der Kodierungspraxis. Konflikte traten sowohl innerhalb einzelner Register als auch zwischen verschiedenen Registern auf. Die folgenden Beispiele verdeutlichen typische Herausforderungen und die daraus abgeleiteten Maßnahmen:

- Unterschiedliche Nutzung des Werts „TX“ (Merkmal *Tumor*):
  Ein Krebsregister hatte technisch nicht die Möglichkeit, den Wert „TX“ zu kodieren, während ein anderes diesen Wert regelmäßig nutzte. Um eine zuverlässige Modellperformance zu gewährleisten, wurde das Training der Machine-Learning-Komponenten entsprechend angepasst.
- Unterschiedliche Detaillierung beim Diagnosecode:
  In einem Register konnte der ICD-10-Diagnosecode technisch nur mit einer Stelle nach dem Dezimalpunkt angegeben werden (z. B. *C41.30 → C41.3* statt *C41.30*). In diesen Fällen wurde - sofern technisch möglich - die fehlende Stelle durch halbautomatisierte Verfahren ergänzt.
- Uneinheitliche Leitlinien zur Kodierung der Stanzenanzahl (Modul Prostata):
  Bei der Erfassung von entnommenen und befallenen Stanzen ergaben sich Inkonsistenzen, insbesondere bei mehrteiligen Stanzen. Hier waren die entsprechenden Leitlinien zur Kodierung nicht eindeutig formuliert, was zu abweichenden Kodierungen für inhaltlich vergleichbare Befunde führte. Nach Identifikation dieses Problems wurden die Leitlinien überarbeitet. Für die weitere Evaluation wurde ein Teil des betroffenen Datensatzes korrigiert.

5.2.3 AP 2.3 Organspezifische Zusatzmodule (15 PM)

Zusätzlich zu den Standardentitäten, die für jede Tumorentität zu kodieren sind, sieht der oBDS für bestimmte Tumorentitäten die Erfassung weiterer, organspezifischer Merkmale vor - die vier sogenannten organspezifischen Zusatzmodule.

Zusatzmodul Prostata

Im Rahmen des Prostata-Zusatzmoduls wurden acht zusätzliche Merkmale zur strukturierten Erfassung entwickelt und verbessert. Einige dieser Felder werden bereits in strukturierter Form an

die Krebsregister übermittelt, sodass hierfür keine Extraktion erforderlich war. Für die verbleibenden Merkmale wurde eine automatische Informationsextraktion aus Freitexten umgesetzt.

Zu Projektbeginn lagen zunächst nur Daten des Krebsregisters Baden-Württemberg vor. Im weiteren Verlauf konnten auch Daten anderer Register einbezogen werden. Die bereits entwickelten Komponenten mussten hierfür nur geringfügig angepasst werden und erzielten auch auf den neuen Daten sehr gute Ergebnisse.

Die Extraktion des Gleason Scores - eines zentralen Merkmals in der Prostatakarzinomdiagnostik - war bereits zuvor technisch möglich, jedoch lagen erst im Projektverlauf geeignete Daten für eine umfassende Evaluation vor. Die Klassifikationsgenauigkeit ist mit F1-Scores von 0,915 (Gesamt), 0,940 (Primary Pattern), 0,929 (Secondary Pattern) und 0,976 (Anlass der Scorebestimmung) als sehr hoch zu bewerten.

Besondere Herausforderungen ergaben sich bei der Extraktion der Angaben zu den Stanzen (z. B. Anzahl befallener bzw. untersuchter Stanzen, maximaler Karzinombefall). Die Entwicklung war aufgrund der vielfältigen Darstellungsformen im Fließtext und uneinheitlich umgesetzter medizinischer Leitlinien besonders aufwändig. Dennoch konnten auch hier, mit F1-Scores von 0,814 (befallene Stanzen), 0,844 (untersuchte Stanzen) und 0,946 (maximaler Karzinombefall) hohe Extraktionsqualitäten erreicht werden.

Darüber hinaus wurde der PSA-Wert (prostataspezifisches Antigen) über die bereits bestehende Komponente zur Laborwerterkennung extrahiert. Die Herausforderung lag dabei insbesondere in der Auswahl des korrekten Wertes bei mehrfacher Nennung innerhalb eines Dokuments, was eine gezielte Kontextanalyse erforderlich machte.

Zusatzmodul Mamma

Für das Mamma-Zusatzmodul standen die benötigten Goldstandards erst im letzten Projektjahr zur Verfügung. Trotz des späten Datenzugangs konnten innerhalb des verbleibenden Zeitraums fünf zusätzliche Merkmale erfolgreich implementiert und evaluiert werden. Teilweise liegen die relevanten Informationen bereits in strukturierter Form vor und sind nicht im Pathologiebefund enthalten, sodass sich die Extraktion auf die verbleibenden Freitextinhalte konzentrierte.

Die Extraktion der Hormonrezeptor-Statusangaben erfolgte über die im Rahmen von HAP 3 entwickelte Biomarkererkennung. Dabei konnten für den Östrogenrezeptor (F1-Score: 0,907), Progesteronrezeptor (0,908) und Her2neu (0,834) sehr gute Ergebnisse erzielt werden.

Die Erkennung der Tumorgröße stellte eine Herausforderung dar. Insbesondere die Unterscheidung zwischen invasivem Karzinom und duktalem Carcinoma in situ (DCIS) gestaltete sich komplex. Dennoch konnten auch hier hohe Extraktionsqualitäten erreicht werden, mit F1-Scores von 0,889 für die Tumorgröße des invasiven Karzinoms und 0,795 für die Größe des DCIS.

Zusatzmodul Malignes Melanom

Auch für das Modul *Malignes Melanom* standen die relevanten Kodierungen erst im letzten Projektjahr zur Verfügung. Innerhalb des verfügbaren Zeitraums wurden vier zusätzliche Merkmale erfolgreich extrahiert und bewertet.

Die Tumordicke war bereits vor Projektbeginn technisch extrahierbar, jedoch lag bislang keine Evaluation der Extraktionsergebnisse vor. Mit Verfügbarkeit geeigneter Daten konnte nun ein F1-Score von 0,945 erreicht werden, was auf eine sehr zuverlässige Erkennung hinweist.

Der LDH-Wert (Laktatdehydrogenase) wird überwiegend strukturiert übermittelt, sodass keine manuelle Annotation zur Bewertung der Extraktion vorliegt. Dennoch wurde die bestehende Komponente zur Laborwerterkennung so erweitert, dass LDH-Werte aus Freitexten nun identifiziert und zurückgegeben werden können - eine Aussage zur Extraktionsgüte ist jedoch nicht möglich.

Ein weiteres zentrales Merkmal ist der minimale Sicherheitsabstand zum Primärtumor. Die Extraktion erwies sich als herausfordernd, da in vielen Fällen mehrere Abstandsangaben vorhanden sind und die korrekte Auswahl eine präzise Kontextanalyse erfordert. Dennoch wurde ein sehr guter F1-Score von 0,922 erzielt.

Für die Erkennung einer Ulzeration - also des Einbruchs des Tumors in tiefere Hautschichten - konnte ein Modell entwickelt werden, das zwischen ulzerierten und nicht-ulzerierten Befunden zuverlässig unterscheidet. Die erreichte Extraktionsqualität mit einem F1-Score von 0,981 belegt die hohe Genauigkeit der Klassifikation.

Zusatzmodul Kolorektales Karzinom

Für das Modul *Kolorektales Karzinom* standen die Daten zu fünf zusätzlichen Merkmalen ebenfalls erst im letzten Projektjahr zur Verfügung. Aufgrund des fortgeschrittenen Projektzeitpunkts konnte jedoch lediglich die Extraktion eines dieser Merkmale - das K-ras-Onkogen - implementiert und evaluiert werden.

Die Extraktion des K-ras-Mutationsstatus erfolgte über die im Rahmen von HAP 3 entwickelte Biomarkererkennung. Das System konnte dieses Merkmal mit hoher Genauigkeit identifizieren und erreichte dabei einen F1-Score von 0,982.

5.2.4 AP 2.4 Zusätzliche Klassifikationen (6 PM)

Über die Standardentitäten und organspezifischen Zusatzmodule hinaus wurden mehrere ergänzende Klassifikationsansätze identifiziert, die potenziell zur weiteren Verbesserung der automatisierten Verarbeitung von Befundtexten beitragen können. Diese Zusatzklassifikationen wurden im Hinblick auf ihre Relevanz priorisiert und - sofern zeitlich und datenseitig möglich - näher untersucht.

Die höchste Priorität hatte die Unterscheidung zwischen Einzel- und Mehrfachbefunden. Für diese Klassifikation standen im letzten Projektjahr Daten aus zwei Krebsregistern zur Verfügung. Auf dieser Basis konnte ein erstes Modell trainiert und eine umfassende Fehleranalyse durchgeführt werden. Dabei zeigte sich, dass sich die Definitionen und Annotationskonventionen zwischen den Registern unterscheiden, was die Modellbildung erschwerte. Entsprechend konnte bislang keine hinreichend zuverlässige Komponente entwickelt werden, die für den Routineeinsatz geeignet wäre. Aufgrund der hohen praktischen Relevanz für die Register ist jedoch eine Weiterverfolgung dieses Themas über das Projektende hinaus angestrebt.

Weitere potenzielle Zusatzklassifikationen - wie die Unterscheidung zwischen operativen und bioptischen Befunden (OP/Stanze), die Identifikation von Metastasen- und Rezidivbefunden sowie die Erkennung nicht meldepflichtiger Diagnosen - wurden im Projektverlauf konzeptionell betrachtet, jedoch nicht implementiert. Ausschlaggebend hierfür waren in erster Linie das Fehlen geeigneter Goldstandards sowie die niedrigere Priorität im Vergleich zu anderen Aufgaben im Projekt, z.B. den organspezifischen Zusatzmodulen. Im Fall der nicht meldepflichtigen Diagnosen bestand darüber hinaus kein gesteigerter Bedarf, da die beteiligten Krebsregister zum Teil bereits über etablierte Mechanismen zur Filterung entsprechender Dokumente verfügen.

5.3 Arbeitspaket 3: Erweiterter Molekularpathologischer Befund (AF 2)

5.3.1 AP 3.1 – 3.4 Erkennung von Genen, Proteinmutationen, Ausprägung genetischer Varianten

Die Ergebnisse für die Teilarbeitspakete AP3.1 – AP3.4 werden im Folgenden als Ganzes dargestellt, da sie eng ineinandergreifen und gemeinsam einen Gesamtprozess für die Auswertung des erweiterten Molekularpathologischen Befundes (AF2) bilden. Die praxisnahe übergreifende Aufgabe, die auch alle in Kapitel 3 genannten Teilziele AF2.1 – AF2.4 umfasst, lautet hierbei wie folgt:

- Molekularpathologische Befunde sollen bzgl. relevanter Gene[1], die durch Krebsregister mittels einer Liste festgelegt werden, systematisch durchsucht werden. Hierbei sollen pro Befund und pro Gen Untersuchungsergebnisse, beurteilende Aussagen und sonstige Informationen bzgl. aufgetretener Gen-Varianten aufgefunden werden.
- Pro Befund und pro befundetem Gen sollen alle relevanten Aussagen interpretiert und zu einem Beurteilungsergebnis zusammengeführt werden. Das Beurteilungsergebnis soll als Klassifikationsentscheidung mit strukturierten Zusatzinformationen ausgegeben werden.
- Richtlinie: Werden Gen-Varianten ausschließlich innerhalb von Tabellen als Rohdaten (z.B. numerische Werte) genannt, so ist dies nicht ausreichend dafür, um eine Klassifikation vornehmen zu können. Stattdessen soll zu der Tabelle ein flankierender Text mit qualifizierenden, beurteilenden Aussagen vorliegen.

Die Lösung der gestellten Aufgabe, d.h. die Erkennung von Aussagen zu untersuchten genetischen Varianten in Befunden und die Ableitung eines Klassifikationsergebnisses mit anschließender strukturierter Ausgabe, geschieht schrittweise für jeden einzelnen Befund gemäß folgender Text-Mining-Prozesskette:

(1) Gene bzw. Aussagen zu Genen auffinden und mit dem umliegenden Kontext extrahieren, im Folgenden als „Gen-Nennung“ bezeichnet (Teilziel AF2.1 / AP 3.1 und 3.4)
(2) Protein-Varianten, cDNA-Varianten, Referenzsequenz-Nummern und Exone aus Gen-Nennungen extrahieren (Teilziel AF2.2 / AP 3.3 und 3.4)
(3) Art der Gen-Nennung feststellen und bzgl. Relevanz für die Ausprägungsbestimmung klassifizieren; irrelevante Gen-Nennungen aussieben (Teilziel AF2.3 / AP 3.2 und 3.4)
(4) Isolierte Klassifikation der Ausprägung für jede relevante Gen-Nennung vornehmen. Danach Zusammenführung aller Einzelklassifikationen zu einer Gesamtklassifikation der Ausprägung pro Gen (Teilziel AF2.4, AP 3.2 und 3.4)
(5) Strukturierte Ausgabe der Klassifikation und der Zusatzinformationen

Im Folgenden wird auf jeden einzelnen Schritt bzgl. des Lösungsansatzes und der qualitativen Zielerreichung eingegangen.

(1) Gene bzw. Aussagen zu Genen auffinden (Teilziel AF2.1, AP 3.1)

Ziel: In diesem Schritt werden alle Informationen inklusive Tabellenzeilen zu einem Gen im Befund aufgefunden. Hierzu werden zunächst Gen-Bezeichnungen detektiert und anschließend mit dem umliegenden Kontext extrahiert. Hierdurch entstehen kleine Einheiten bestehend aus Paaren

[1] „Gen“ steht im Folgenden stellvertretend für die Gen-Bezeichnungen in HGNC-Schreibweise, welche jedoch zum Teil nicht ohne Kontext von dem jeweiligen Protein zu unterscheiden ist. Für eine bessere Lesbarkeit des Berichts wird in diesem Abschnitt nur von Genen, Gen-Bezeichnungen oder Gen-Nennungen gesprochen.

<Gen-Bezeichnung, Kontext> für die Weiterverarbeitung, die im Folgenden „Gen-Nennung“ genannt werden.

Ansatz: Die Erkennung der relevanten Gene geschieht grundsätzlich durch zeichenbasiertes Matching auf Basis der vom HUGO Gene Nomenclature Committee (HGNC) festgelegten Schreibweise des jeweiligen Gens. Die aktuell für KLR-Baden-Württemberg relevanten Gene werden mithilfe eines Wortschatzes (Gen-Liste) manuell gepflegt. Die Gen-Liste kann jederzeit ergänzt werden und enthält derzeit ca. 90 Gene. Die Gen-Liste ist eine tabellarische Auflistung der Gen-Bezeichnungen, sie wird derzeit als Datei im Format Comma-Separated Values (CSV; durch Trennzeichen getrennte Werte) geführt und beim Start des Programms als Input eingelesen. Die Verwendung einer CSV-Datei (o.Ä.) als programmexterne Ressource hat den Vorteil, dass die Gen-Liste transparent ist und jederzeit leicht fortgeschrieben werden kann, wenn neue zu extrahierende Gene / Biomarker hinzugenommen werden sollen. Die Gen-Liste kann durch Mitarbeitende der Krebsregister zukünftig selbstständig gepflegt und ergänzt werden.

Gen-Schreibweisen: Eine umfassende Untersuchung von Gen-Bezeichnungen in einem Korpus aus 21.000 Meldungen hat ergeben, dass Gene nicht nur in der präferierten standardisierten HGNC-Schreibweise auftreten, sondern dass auch alternative Schreibweisen verwendet werden. Regelmäßig werden z.B. Bindestriche oder Leerzeichen vor Ziffern eingefügt (z.B. „IDH1“ => „IDH-1“ oder „IDH 1“) oder bestimmte einzelne Buchstaben kleingeschrieben (z.B. „PDGRFA“ => „PDGFRa“ oder „PIK3CA“ => „Pik3ca“, „PiK3CA“, „PIK3Ca“). Die Groß- und Kleinschreibung darf bei der Identifikation von Genen in den Texten allerdings nicht generell ignoriert werden, da es hierdurch zu einer unerwünschten Verwechslung mit Proteinen oder mit anderen Wörtern (z.B. das Gen „KIT“ mit einem „test kit“) kommen kann. Stattdessen müssen Gene „case-sensitiv“ (also als exakte Treffer aller Zeichen mit vorgegebener Groß- und Kleinschreibung) identifiziert werden. Die wenigen von der HGNC-Standardschreibweise abweichenden tolerierten Schreibweisen werden als „White-List“ in der „Gen-Liste“ mitgepflegt und dort der Standardbezeichnung als Alias-Bezeichnung zugeordnet.

Als Ergebnis des Verarbeitungsschrittes resultiert pro Befund eine Menge von im Befund genannten Genen. Pro Gen resultiert wiederum eine Liste aller konkreten Gen-Nennungen, d.h. jeweils Paaren <Gen, Kontext>.

Qualität / Zielerreichung: Im Testkorpus aus 21.000 Befunden konnten insgesamt ca. 700.000 Gen-Nennungen für ca. 90 unterschiedliche Gene detektiert werden. Nachdem Gen-Nennungen für dasselbe Gen pro Befund zusammengeführt wurden, resultierten ca. 320.000 zu bewertende Gene. Mit den implementierten Pattern-Matching-Regeln sind für den Testkorpus keine False-Positive-Fälle festgestellt geworden. False-Negatives (i.e. nicht erkannte Gene) treten nur in seltenen Ausnahmefällen auf: Durch die iterative und systematische Aufnahme regelhafter Alias-Bezeichnungen im Verlauf des TeMeK-Vorhabens wurde die Hauptfehlerquelle eliminiert. In den sehr seltenen verbleibenden Fällen, in denen ein Gen nicht erkannt wurde, lag meist ein willkürlicher Tippfehler vor. Insgesamt stellen Fehler bei der Gen-Erkennung somit kein Problem dar. Der Prozessschritt konnte im Projektzeitraum mithilfe der Gen-Liste und zeichenbasierten Erkennung transparent, effektiv und effizient gelöst werden. Er ist zudem mit Blick auf AP 3.4 dahingehend flexibel gelöst, dass zukünftig neu zu untersuchende Gene durch das Register-Personal einfach in die Gen-Liste aufgenommen werden können (inklusive regelhafter Alias-Bezeichnungen), ohne dass Algorithmen geändert werden müssen.

(2) Protein-Varianten, cDNA-Varianten, Referenzsequenz-Nummern und Exone aus Gen-Nennungen extrahieren (Teilziel AF2.2 / AP 3.3)

Ziel: In diesem Schritt werden die in der Gen-Nennung (<Gen, Kontext>) nach der Nomenklatur der Human Genome Variation Society (HGVS)[2] angegebenen Protein- und cDNA-Varianten sowie Referenzsequenz-Nummern extrahiert. Des Weiteren werden die in der Gen-Nennung angegebenen Exone mit ihren Nummern extrahiert. Häufig werden mehrere Exone gleichzeitig untersucht und in unterschiedlichen Mengenschreibweisen wie z.B. „1-5“, „1 -- 5“, „1 bis 5“ angegeben. Daher besteht ein weiterer Schritt darin, die Mengenschreibweisen in eine einheitliche Form mit Bindestrich und ohne Leerzeichen zu überführen, z.B. „1-5“. Die Extraktion und Normalisierung erfolgt zu dem Zweck, dass die entsprechenden Varianten, Nummern und Exone in der strukturierten Ausgabe zu einem Gen als Zusatzinformation ausgegeben werden können.

Ansatz: Die Erkennung von Protein- und cDNA-Varianten sowie Referenzsequenz-Nummern erfolgt mithilfe von regulären Ausdrücken. Diese sind transparent, effektiv und effizient. Um die regulären Ausdrücke definieren zu können, wurde zum einen der bereits genannte Korpus von 21.000 Meldungsdokumenten auf Vorkommen von HGVS-Notationen untersucht. Zum anderen wurden die Bildungsregeln der HGVS-Nomenklatur zur Rate gezogen. Pro Gen-Nennung (<Gen, Kontext>) resultiert für die Protein-Varianten, cDNA-Varianten und Referenzsequenz-Nummern je eine Liste von entsprechenden Originalstrings aus dem Befundtext, z.B.:

- Protein-Varianten: *p.Ser106fs*, p.Asp445Tyr*
- cDNA-Varianten: *c.375+1G>T, c.403dupT*
- Referenzsequenz-Nummern: *NM_000546.5*

Die Erkennung von Exonen erfolgt ebenfalls mithilfe von regulären Ausdrücken. Um die regulären Ausdrücke definieren zu können, wurde derselbe Korpus (21.000 Meldungsdokumente) auf das Vorkommen von Informationen zu Exonen untersucht. Im Anschluss an die Extraktion der Originalstrings aus dem Befundtext werden diese mithilfe einer Abfolge von Standardisierungsschritten in eine einheitliche Schreibweise gebracht. Pro Gen-Nennung resultiert eine Liste von Nummern, die in einer standardisierten Schreibweise ausgegeben werden, z.B. „1, 3, 5-8, 14“.

Qualität / Zielerreichung: False-Negatives sind im Kontext dieses Prozessschrittes Wörter, welche p-/cDNA-Varianten, Referenzsequenz-Nummern oder Exon-Angaben repräsentieren, aber nicht als solche erkannt werden. Da Pattern sowohl ausgehend vom Korpus (d.h. dem Auftreten in den Realdaten) als auch ausgehend von der deklarativen HGVS-Nomenklatur abgeleitet wurden, besteht eine sehr gute Abdeckung. Für cDNA-Varianten, Exone sowie Referenzsequenz-Nummern sind im 21.000-Korpus keine False-Negative-Fälle bekannt geworden. Für Protein-Varianten wurden in sehr seltenen Einzelfällen Pattern nicht erkannt, die jedoch keinen gängigen Bildungsregeln folgten und daher vermutlich Tippfehler darstellten. False-Positives sind im Kontext dieses Prozessschrittes Wörter bzw. Ziffernfolgen, welche keine p-/cDNA-Varianten, Exon-Angaben oder Referenzsequenz-Nummern darstellen, aber fälschlicherweise als solche interpretiert werden. Für den Testkorpus sind keine False-Positives bekannt geworden. Dieser Schritt ist mithilfe der regulären Ausdrücke und zeichenbasierten Erkennung somit transparent, effektiv und effizient gelöst. Er ist zudem in Bezug auf AP 3.4 dahingehend nachhaltig gelöst, dass zukünftig neu zu untersuchende Gene und neu auftretende p-/cDNA-Varianten denselben

[2] https://hgvs-nomenclature.org/stable/

Bildungsregeln nach HGVS-Nomenklatur folgen und somit ebenfalls erkannt werden, ohne dass Algorithmen geändert werden müssen.

(3) Art der Gen-Nennung feststellen, Relevanz klassifizieren (Teilziel AF2.3 / AP 3.2, 3.4)

Ziel: Nicht jede Gen-Nennung eines Befunds ist für die Klassifikation der Ausprägung relevant. Hochrelevant ist Fließtext, der beurteilende Aussagen enthält. Im Gegensatz dazu sind allgemeingültige Aussagen zu Gen-Mutationen irrelevant und dürfen nicht versehentlich als detektierte Gen-Mutation interpretiert werden. In diesem Schritt wird daher die Art der Gen-Nennung festgestellt und als „Relevanz-Klasse“ an die Gen-Nennung angeheftet.

Ansatz: Es wird zwischen drei Relevanz-Klassen unterschieden, die sequenziell auf Basis von Schlüsselwortkombinationen und zeichenbasierten Pattern abgeprüft werden. Pro Gen-Nennung <Gen, Kontext> resultiert eine der folgenden drei Relevanz-Klassen:

- Irrelevanz: Dies sind Gen-Nennungen in Form von Fließtext, wie Untersuchungsankündigungen, Untersuchungsmethoden oder allgemeingültige Aussagen, aus denen keine Ausprägung abgeleitet werden darf. Beispiele:
    - „*Aktuell rekrutieren klinische Studien Patienten mit fortgeschrittenen soliden Tumoren u.a. NSCLC, die eine KRAS p.G12C Mutation aufweisen.*“
    - „*In der wissenschaftlichen Literatur zeigten Patienten mit gemeinsam auftretenden deletären STK11- und aktivierenden KRAS- Mutationen ein schlechteres Ansprechen auf Chemotherapie (PMID: 29191602).*“
- Tabelle: Dies sind Gen-Nennungen, die innerhalb von Tabellen vorkommen. Eine ausschließliche Nennung von Gen-Varianten innerhalb von Tabellen als Rohdaten soll nicht ausreichend dafür sein, um eine Ausprägungs-Klassifikation vornehmen zu können. Stattdessen soll zu der Tabelle ein flankierender Text mit beurteilenden Aussagen vorliegen. Beispiel:
    - *Gen|Variante (Protein)|Variante (cDNA)|RefSeq-ID|ID*
      ERBB2|p.Pro1170Ala|c.3508C>G|NM_004448.3|rs1058808
- Beurteilung: Dies sind Gen-Nennungen in Form von Fließtext, die beurteilende Aussagen zum Vorliegen einer Gen-Mutation enthalten. Auf Basis dieser Klasse wird im nächsten Verarbeitungsschritt die Ausprägung eines Gens bestimmt. Beispiele:
    - *„Es zeigt sich eine aktivierende KRAS Exon 2 Variante (p.G12R).“*
    - *„Eindeutiger Nachweis einer deletären BRCA2-Mutation, BRCA2:p.Gln1429*.“*

Qualität / Zielerreichung: Von Interesse für die Beurteilung der Qualität der Klassifikation sind folgende zwei Fehler-Klassen. False-Negatives: Gen-Nennungen werden als nicht relevant (Irrelevanz oder Tabelle) klassifiziert, obwohl diese eine Beurteilung enthalten. Eine Beurteilung für das Gen wird somit übersehen. False-Positives: Gen-Nennungen werden als relevant klassifiziert, obwohl diese keine spezifischen Beurteilungen für die behandelte Person, sondern z.B. allgemeingültige Aussagen zu genetischen Varianten enthalten. Als Folgefehler wird ggf. fälschlicherweise eine Gen-Variante als nachgewiesen dokumentiert. Die Eingrenzung und Abschätzung möglicher Fehler wurde durch eine statistische Analyse des Korpus mit 21.000 Meldungen und 320.000 zu bewertenden Genen (als Ergebnis von Schritt 1) durchgeführt:

- Von den 320.000 Genen werden ca. 12.000 ausschließlich mit Immunhistochemie berichtet. Diese werden in der Klassifikations-Pipeline von Averbis (AP 3.5) verarbeitet und daher in diesem Teilschritt ignoriert. Von den 320.000 Genen werden ca. 212.000

ausschließlich in tabellarischen Aufzählungen bzw. Tabellen genannt, d.h. ohne Fließtext. Von den 320.000 Genen resultieren damit ca. 96.000 Gene, die über mindestens eine Fließtext-Gen-Nennung verfügen. Von diesen 96.000 Genen mit Fließtext gibt es wiederum ca. 62.000 Gene, die als „relevant“ klassifiziert wurden. Die Einträge wurden stichprobenartig dahingehend untersucht, ob allgemeine Aussagen fälschlicherweise als Beurteilung erkannt wurden (mögliche False-Positives). Solche Fälle konnten nicht aufgefunden werden. Es wird daher davon ausgegangen, dass die False-Positive-Rate bei der Relevanz-Klassifikation sehr gering ist und unter 0,5% liegt.

- Von den 96.000 Genen mit Fließtest gibt es im Umkehrschluss ca. 34.000 Gene, die als „irrelevant“ klassifiziert wurden. Von diesen 34.000 Genen haben ca. 16.000 nur eine einzige Gen-Nennung im Befund. Aufgrund von Stichproben gehen wir mit einem sehr geringen Fehler von unter 0,5% (d.h. < 100 Gene) davon aus, dass tatsächlich keine Beurteilung berichtet wird. Bei den anderen 18.000 Genen mit mindestens zwei Gen-Nennungen im Befund schätzen wir denselben Fehler auf ca. 1 – 1,5% ein, d.h. dass hier ca. 180 – 360 eigentlich beschriebene Beurteilungen nicht erkannt werden. Im Mittel gehen wir somit bei den 34.000 als „irrelevant“ klassifizierten Genen von höchstens 350 False-Negatives aus, was einer False-Negative-Rate von 1% entspricht.

Insgesamt lässt sich mit den eingesetzten Text-Mining-Techniken (Schlüsselwörter, zeichenbasierte Pattern) die Relevanz-Klassifikation, d.h. die binäre Entscheidung, ob für ein im Befund genanntes Gen tatsächlich eine relevante beurteilende Aussage vorliegt und das Gen daher zu dokumentieren ist, sehr gut lösen, bei nur geringen False-Positive- (< 0,5%) und False-Negative-Raten (ca. 1%). Die Relevanz-Klassifikation ist zudem dahingehend allgemein gelöst, dass mit Blick auf AP 3.4 keine Veränderungen für die Erkennung relevanter Gen-Nennungen notwendig werden, wenn neue Gene und Varianten auftreten und in Befunden erkannt werden sollen.

(4) Klassifikation der Ausprägung pro Gen (Teilziel AF2.4, AP 3.2 und 3.4)

Ziel: Jede vorher (Ergebnis von Schritt 3) als „relevant“ klassifizierte Gen-Nennung <Gen, Kontext> wird in diesem Schritt zunächst isoliert betrachtet, d.h. unabhängig von anderen potenziellen Gen-Nennungen derselben Gen-Art im Befund ausgewertet. Basierend auf dem Kontext muss für die Gen-Nennung im Wesentlichen eine Klassifikation bzgl. der drei Hauptklassen „Mutation (M)“, „nicht bestimmbar (N)“ und „Wildtyp (W)“ vorgenommen werden. Daneben gibt es weitere Nebenklassen für Spezialfälle („Sonstiges (S)“, „unbekannt (U)“). Im Anschluss an die isolierten Klassifikationen aller einzelnen Aussagen zum Gen im Befund werden diese zu einem Gesamtergebnis pro Gen und Befund zusammengeführt.

Ansatz: Die Klassifikation basiert auf einem intuitiven regel- und schlüsselwortbasierten Ansatz, welcher aussagenlogische Werte aus dem Satz ableitet und kombiniert. Auf Basis der erkannten Schlüsselwörter für die genannte Variante wird zunächst ein logischer Grundwert festgelegt, z.B. folgt aus Schlüsselwörtern für Mutationen wie „*Fusion*“ / „*Deletion*“ / „*Translokation*“ etc. der Grundwert „Mutation positiv“ oder aus dem Schlüsselwort „*Wildtyp*“ der Grundwert „Mutation negativ“. Analog dazu werden logische Werte aus erkannten Beurteilungsphrasen gezogen, z.B. folgt aus bestätigenden Schlüsselphrasen wie „*findet sich*“ / „*Nachweis*“ / „*detektiert*“ der Wert „Mutation positiv“, hingegen folgt aus Schlüsselphrasen wie „*Hinweis*“ / „*Verdacht auf*“ / „*Anhalt*“ der Wert „Mutation unsicher“. Schließlich werden weitere logische Operatoren im Kontext erkannt, z.B. führen Schlüsselwörter wie „*kein*“ / „*fehlen*“ / „*ohne*“ zu einer Negation des logischen Grundwertes. Durch Kombination der logischen Werte und Operatoren auf Basis eines entwickelten Regelwerkes ergibt sich die Ausprägungs-Klasse. Beispiele:

- *„Es **zeigt sich** eine Nukleotidsubstitution von G>A an Position c.395 des IDH1 Gens, entsprechend einer p.R132H **Mutation**."* => Variante: positiv, Beurteilung: positiv, Negation: nein => Klasse: M (Mutation)
- *„In dem Plattenepithelkarzinom **liegt keine Amplifikation** von MET vor."* => Variante: positiv, Beurteilung: positiv, Negation: ja => Klasse: W (Wildtyp)
- *„Die Analyse der Coverage-Daten ergab **Hinweise** auf das mögliche Vorliegen einer **Amplifikation** des FGFR1-Genlokus."* => Variante: positiv, Beurteilung: unsicher, Negation: nein => Klasse: N (nicht bestimmbar)

Die auf diese Weise erhaltenen Einzelklassifikationen der zu einem Gen im Befund detektierten Aussagen werden miteinander verglichen. Liegt ein Konsens vor, so wird die entsprechende Klasse (M, N, W) zurückgegeben, bei widersprüchlichen Aussagen wird die Warnung „Konflikt" erzeugt.

Qualität / Zielerreichung: Analog zu Schritt 3 können Fehl-Klassifikationen auftreten, i.e. False-Negatives (FN), wenn ein Gen fälschlicherweise als „Wildtyp" klassifiziert wird, obwohl eine Mutation vorliegt, und False-Positives (FP), wenn ein Gen fälschlicherweise als „Mutation" klassifiziert wird, obwohl ein Wildtyp vorliegt. Die Güte des umgesetzten Klassifikationsverfahrens für die Ausprägung wurde auf dem bereits genannten Korpus von 21.000 Meldungen bzgl. der Art, Ursache und Häufigkeit von Fehlern analysiert.

Problem 1 - Wortschatz: Da die Klassifikation auf der Erkennung von Schlüsselphrasen basiert, liegt eine mögliche Fehlerquelle in einem noch unausgereiften Wortschatz (Nicht-Eignung oder Fehlen von Schlüsselphrasen). Der Wortschatz konnte im Projektverlauf iterativ verfeinert werden und stellt mit Projektabschluss keine nennenswerte Fehlerquelle mehr da.

Problem 2 - schwierige Negation: Eine zweite mögliche Fehlerquelle liegt in der Kombination der logischen Werte und hierbei vor allem in falsch erkannter bzw. falsch zugeordneter Negation, z.B.:

- *„Die **detektierte** c-KIT Exon11 **Mutation** ist in den Datenbanken **nicht** zu **finden**."*

Im Beispiel muss durch den Algorithmus erkannt werden, dass sich die Negation auf die „*Datenbank"* und nicht auf die „*detektierte Mutation"* bezieht. Die korrekte Klassifikation ist Mutation, nicht Wildtyp.

Problem 3 – Vermischung von Aussagen: Eine dritte erkannte Fehlerquelle bzw. Herausforderung liegt in der Vermischung von Aussagen zu unterschiedlichen Varianten desselben oder auch verschiedener Gene innerhalb derselben Satzkonstruktion, z.B.:

- *„Die Tumorzellen […] zeigen eine PIK3CA **Mutation** (p.E545K) in **Exon 10** sowie **Wildtyp**sequenzen in **Exon 8 und 21**."*
- *„In der molekularpathologischen Mutationsanalyse findet sich eine relevante **Mutation** des **KRAS**-Gens bei **Wildtyp** des **NRAS**-Gens."*

In den gezeigten Beispielen muss analog zu der Negationsproblematik der Kontext, der zur Bewertung der Aussage zu einem Gen bzw. Exon-Mutation verwendet wird, sorgfältig zugeordnet werden.

Problem 4 – unterschiedliche Varianten: Bei der Zusammenführung der Klassifikationen für die einzelnen Gen-Nennungen eines Gens treten im aktuell implementierten Ansatz vermeintlich konfligierende Aussagen dadurch auf, dass unterschiedliche Varianten desselben Gens beurteilt werden, z.B.:

- Aussage A: *„Kein Nachweis von MET **Exon 14 Skipping**.“* => Wildtyp
- Aussage B: *„Im vorliegenden Gewebe Nachweis einer vermutlich aktivierenden MET **Mutation: p.Leu1195Phe** (NM_000245.4).“* => Mutation

Dieses Problem kann dadurch ausgeräumt werden, dass die Aggregationsebene der Zusammenführung der Einzelklassifikationen verfeinert wird: Diese darf nicht auf der Ebene von Genen, sondern muss auf der von Gen-Varianten stattfinden.

Die genannten Fehler und Herausforderungen traten im Textkorpus mit folgenden Häufigkeiten auf: Von den 62.000 in Prozessschritt 3 als relevant erkannten Genen, also solchen, für die eine Ausprägung dokumentiert werden muss, wurden ca. zwei Drittel (40.000) atomar berichtet. Mit „atomar“ ist hierbei gemeint, dass keine Aussagen zu unterschiedlichen Varianten oder Genen in derselben Satzkonstruktion zusammengefasst wurden, so dass hier Problem 3 als Fehlerquelle ausgeschlossen werden kann.

Innerhalber dieser Teilmenge trat mit einer Häufigkeit von ca. 0,5% falsch zugeordnete Negation (Problem 2) als Ursache für Fehlklassifikationen auf. Der Algorithmus kann an dieser Stelle weiter verbessert werden, indem der Kontext der Gen-Nennung reduziert bzw. die verwendeten grammatikalischen Strukturen (Parse-Tree) mit höherer Detailtiefe berechnet werden.

Innerhalb derselbe Teilmenge kam es zudem bei 2% der Gene zur Ausgabe „Konflikt“, so dass keine Gesamtausprägung bestimmt werden konnte. Die Analyse der Ursache ergab folgendes Bild: Bei ca. 30% (entspricht 0,6% der 40.000 Gene) lag ein tatsächlicher schwacher Widerspruch der Aussagen im Befund zwischen den Klassen M (klare Mutation) und N (Hinweis auf Mutation) vor, u.a.:

- Aussage A: *„In der Mutationsanalyse zeigten sich **Hinweise auf** […] **Varianten** […] in folgenden Genen: TP53:p.Arg273Cys.“* => Klasse N
- Aussage B: *„Auch die hier **detektierten Mutationen** in TP53 und SMAD4 kamen in der vorigen Analyse […] zur Darstellung.“* => Klasse M

Diese Konfliktart ließe sich zukünftig dadurch auflösen, indem bei der Zusammenführung von Aussagen der Klasse N gegenüber M generell Vorrang eingeräumt wird.

Bei weiteren ca. 50% (entspricht 1% der 40.000 Gene) lag Problem 4 vor, d.h. hier wurde zu unterschiedlichen Varianten desselben Gens mit unterschiedlichem Ergebnis (Mutation versus Wildtyp) berichtet. Dieses Problem lässt sich leicht auflösen, indem die Aussagen auf der Ebene der Gen-Varianten und nicht Gen-Art aggregiert werden.

Bei weiteren ca. 20% (entspricht 0,4% der 40.000 Gene) war eine Fehlklassifikation einer der zusammengeführten Gen-Nennungen zu einem Gen die Ursache für den Konflikt.

Für ca. ein Drittel (22.000) der 62.000 zu dokumentierenden Gene lag grundsätzlich Problem 3 vor, d.h. es findet eine Beurteilung von mehreren Genen innerhalb desselben Satzgefüges statt. Diese Aussagen bergen ein höheres Risiko dafür, dass Ausprägungen falsch zu den genannten Genen zugeordnet werden und Fehlklassifikation resultieren. Da der Anteil dieser vermischten Aussagen sehr groß und somit kein Randphänomen ist, wurde er genauer untersucht, um mögliche Fehler abschätzen und den Text-Mining-Ansatz für unterschiedliche Fallgruppen ggf. gezielt weiterentwickeln zu können:

Fall 1 – Hauptsatz mit einheitlicher Aussage und Aufzählung: Die Vermischung ist unproblematisch, wenn es sich um eine einheitliche Aussage handelt, bei der die Beurteilung für

alle in einer Aufzählung genannten Gene gleichzeitig getroffen wird. Dies sind häufig universelle Wildtyp-Beurteilungen, wie z.B.:

- *„Alle weiteren Gene, bzw. Sequenzregionen (insb. die angefragten Gene: **ALK, KRAS, FGFR1, MET, NRAS, PIK3CA, ROS1**) liegen lt. o.g. Grenzwerten für die Auswertung des nNGM Konsortiumspanels als Wildtyp vor.“* => Klasse Wildtyp für alle Gene
- *„In der Fusionsanalyse konnte kein spezifisches Fusionstranskript **(ALK, ROS1, RET, NTRK1-3)** nachgewiesen werden.“* => Klasse Wildtyp für alle Gene

Solche Fälle können algorithmisch daran erkannt werden, dass lediglich ein einziger Varianten-Name bzw. explizit „Wildtyp“ aufgeführt ist UND eine Aufzählung vorliegt. Im Fall 1 kann daher die Ausprägung zunächst für den Satz als Ganzes klassifiziert werden und im Anschluss jedem der in der Aufzählung genannten Gene einzeln zugewiesen werden.

Fall 2 – Aufzählung von je vollständigen Einzelaussagen: Das Satzgefüge beinhaltet eine Aufzählung von je vollständigen Einzelaussagen, die jeweils aus einer Beurteilungsphrase sowie Varianten-Name/Wildtyp bestehen, z.B.:

- *„Nachweis einer aktivierenden **KRAS**-Mutation, kein Nachweis einer **EGFR**-Mutation, kein Nachweis einer **BRAF**-Mutation.“*

Der Lösungsansatz besteht darin, dass das Satzgefüge bzw. die Aufzählung aufgelöst und dann jeder Anteil einzeln klassifiziert wird. Im Gegensatz zu Fall erfolgen Zerlegung und Klassifikation somit in umgekehrter Reihenfolge.

Fall 3 – Hauptsatz mit vermischten Aussagen: In Vergleich zu Fall 2 kann der Satz nicht in vollständige Einzelaussagen zerlegt werden, da es nur eine Beurteilungsphrase gibt, die sich gleichzeitig auf mehrere Varianten/Wildtyp bezieht, z.B.:

- *„In der molekularpathologischen Mutationsanalyse findet sich eine relevante Mutation des **KRAS**-Gens bei Wildtyp des **NRAS**-Gens.“*
- *„Es zeigen sich **BRAF** und **EGFR** Wildtypsequenzen sowie eine **KRAS** Exon 2 Mutation (p.G12D).“*

Solche Fälle sind schwieriger zu lösen, es könnte eine Trennung der Aussagen durch Nutzung detaillierterer Parse-Trees erfolgen. Insgesamt treten die gut zu erkennenden und aufzulösenden Fälle 1 und 2 in den 22.000 vermischten Gen-Beurteilungen mit einer sehr großen Mehrheit von über 95% auf, d.h. Fall 3 mit ca. 5%. Bezogen auf die Gesamtzahl der vorzunehmenden Klassifikationen (62.000) bildet Fall 3 einen Anteil von ca. 1,8%. Da er somit relativ selten auftritt, könnte man in der Praxis vorerst mit einer Warnung und anschließenden manuellen Auflösung durch die Dokumentierenden leben, bis eine Verfeinerung des Algorithmus gezielt für diesen Fall stattgefunden hat.

Insgesamt kann festgehalten werden, dass die Klassifikation der Ausprägung mit den beschriebenen Text-Mining-Verfahren sehr gut lösbar ist. Der vorgestellte Ansatz kann im aktuell implementierten Stand die „unvermischten“ Aussagen (ca. 65% aller Fälle) mit einem Fehler von ca. 0,5% und mit Konflikten bei ca. 2% (Bedarf für manuelle Auflösung) bereits jetzt sehr gut lösen. Die Konflikte können hierbei mit einer nur kleinen Anpassung des Regelwerks auf ca. 0,4% weiter reduziert werden. Diese Werte sind auf die „vermischten“ Aussagen des Falls 1 (einheitliche Aussage für Gen-Aufzählungen) übertragbar. Im Fall 2 und 3 nimmt der aktuelle Implementierungsstand noch keine Differenzierung zwischen den unterschiedlichen Beurteilungen der Gene im selben Satz vor, so dass es hier derzeit gehäuft zu Fehlern kommen könnte. Hier wird

daher eine Warnung ausgegeben. Wie dargelegt kann der Ansatz jedoch mit kleinem Aufwand erweitert werden, so dass die Fälle weitestgehend aufgelöst werden können.

Die Ausprägungsklassifikation ist generisch gelöst und wird für alle Gene und Gen-Varianten in derselben Form angewendet. In Bezug auf AP 3.4 werden somit keine Änderungen notwendig, um neue Gene und Varianten in Befunden erkennen zu können, abgesehen von der Gen-Liste (siehe Schritt 1). Falls zukünftig weitere Varianten-Arten dokumentiert werden sollen, die bisher noch nicht extrahiert werden (z.B. Keimbahnmutationen), so kann dies durch eine einfache Ergänzung des transparent vorliegenden Wortschatzes erreicht werden.

(5) Strukturierte Ausgabe der Klassifikation und der Zusatzinformationen

Ziel: Im letzten Verarbeitungsschritt werden alle extrahierten Informationen sowie erzielten Klassifikationen ausgegeben. Die Ausgabe besteht aus einem strukturierten Teil, der als direkter Input für die strukturierte Dokumentation gemäß Plattform § 65c[3] gedacht ist, sowie aus zusammenfassendem, farblich markiertem Fließtext, der den Dokumentierenden die Verifikation der Ergebnisse erleichtern soll.

Ansatz und Qualität / Zielerreichung: Die in den vorangegangenen Schritten extrahierten Informationen pro Befund und Gen werden in die entsprechenden strukturierten Felder ausgegeben. Außerdem werden Warnungen erzeugt, um auf Unsicherheiten (problematische Negation, Vermischung von Aussagen) und Konflikte hinzuweisen. Bei der strukturierten Ausgabe kommt es zu keinen neuen Fehlern. Wie jedoch bereits in Schritt 4, Problem 4 erläutert, muss zukünftig noch eine feinere Aggregationsebene bei der Ausgabe gewählt werden: Klassifizierte Ausprägungen für Gen-Nennungen müssen pro Gen auf der Ebene der Varianten zusammengeführt werden, nicht auf der Ebene der (standardisierten) Gen-Namen, um zwischen unterschiedlichen Ergebnissen (Mutation / Wildtyp) bei mehreren untersuchten Exonen oder Varianten-Arten differenzieren zu können.

Bei der Angabe des zusammenfassenden Fließtextes gilt das Prinzip, dass (1.) alle Gen-Nennungen inklusive ihres berücksichtigten Kontextes aufgeführt werden, dass (2.) die für die Klassifikationsentscheidung maßgeblichen detektierten Schlüsselphrasen farblich entsprechend eines Farbcodes markiert sind und dass (3.) die für die Einbeziehung einer Gen-Nennung in die Ausprägungsklassifikation maßgebliche Relevanz-Klasse mittels einer entsprechenden Überschrift erkennbar ist. Auf diese Weise kann schnell nachvollzogen werden, wie das Ergebnis entstanden ist, und – falls Klassifikationen noch fehlerhaft sind – welcher Verarbeitungsschritt ursächlich war, um das Verfahren ggf. iterativ zu verbessern. Die folgende Tabelle und Abbildung zeigen eine beispielhafte strukturierte Ausgabe sowie den zusammenfassenden Fließtext für das Gen „BAP1“:

Gen: BAP1	*Protein-Variante:* p.His94Arg
Name der Variante: Mutation	*cDNA-Variante:* c.281A>G
Klassifikation / Ausprägung: M	*Exone:* -

[3]https://plattform65c.atlassian.net/wiki/spaces/Dokumentat/pages/86311017/Genetische+Varianten

Beurteilung / Ergebnis:

Die Mutationsanalyse zeigte eine Mutation in BAP1 (p.His94Arg), lokalisiert in der N-terminalen Peptidase C-12 Domäne.

Tabellen:

Gen|Mutation|cDNA|Allelfrequenz
BAP1|p.His94Arg|c.281A>G|48.699

Gen|Gen|Gen|Gen|Gen|Gen|Gen|Gen
AURKA|AURKB|AXIN1|AXIN2|AXL|B2M|BAP1|BARD1

Sonstige Informationen:

Mutationen in BAP1 sind rekurrente Events in Mesotheliomen.

5.3.2 Arbeitspaket 3.5 Verbindung mit HAP2

Bereits vor der Bereitstellung der Daten zu diesem Arbeitspaket wurden erste Analysen durchgeführt, um relevante Biomarker im Kontext der Tumordokumentation zu identifizieren und deren Repräsentation in den für HAP 2 vorgesehenen Textdaten zu bewerten. Auf Basis dieser Vorarbeiten wurde eine prototypische Extraktionskomponente entwickelt, die jedoch nicht in den regulären Verarbeitungsprozess integriert wurde.

Erst nach Sichtung eines neu bereitgestellten Datensatzes und gemeinsamer Abstimmung mit FKIE und dem Krebsregister Baden-Württemberg (KR BW) wurde entschieden, dass Averbis sich vorrangig auf die Extraktion immunhistochemischer Informationen konzentriert, während genetische Merkmale primär durch das FKIE bearbeitet wurden. Dies diente auch dazu, Parallelentwicklungen zu vermeiden.

Da der bereitgestellte Goldstandard nur Informationen zu acht Biomarkern enthielt und die Textausschnitte zum Teil zu kurz waren bzw. keine Bewertung enthielten, wurde zusätzlich ein eigener Goldstandard erstellt, um die Modellqualität gezielter evaluieren und weiterentwickeln zu können.

Anhand des zur Verfügung gestellten Goldstandards konnte eine Auswertung erfolgen, die sich auf die korrekte Erkennung von Nennungen verschiedener Biomarker konzentrierte. Für acht der ausgewerteten Marker wurden F1-Scores zwischen 0,962 und 1,000 erreicht, bei durchgängig perfekter Precision (1,000). Ein auffälliger Ausreißer war der Marker p16 mit einem F1-Score von 0,886. Ursache war hier häufig eine Kodierung von p16, obwohl im Freitext lediglich ein Hinweis auf Positivität bzw. Negativität für humane Papillomviren (HPV) enthalten war.

Zur erweiterten Validierung wurde ein interner Goldstandard mit insgesamt 20 Biomarkern sowie zur Mikrosatelliteninstabilität (MSI) erstellt. In dieser Evaluation lagen die F1-Scores zwischen 0,653 und 0,957, wobei MSI bei 0,744 lag. Reduziert man die Auswertung auf jene Biomarker, die auch im bereitgestellten Goldstandard enthalten waren, ergaben sich F1-Scores zwischen 0,736 und 0,939, was die Vergleichbarkeit und Reproduzierbarkeit der Ergebnisse bestätigten. Variierende F1-Werte lassen sich zudem auf die zum Teil geringe Anzahl verfügbarer Trainingsbeispiele zurückführen.

Die Beurteilung der extrahierten Biomarker erfolgt kontextbasiert. Neben der reinen Erkennung, die zeichenbasiert erfolgt, wird auch deren Status (positiv, negativ, other) klassifiziert. Bei

mehrfacher Nennung und Beurteilung eines Markers innerhalb eines Dokumentes muss zudem die aktuelle und damit relevante Beurteilung ausgewählt werden.

Ergänzend wurde die Erkennung numerischer Angaben, wie etwa beim Ki-67-Proliferationsindex, implementiert, um quantitative Befundinhalte strukturiert verfügbar zu machen.

Zur Verbesserung der Nutzbarkeit der Komponente wurde zudem eine strukturierte und übersichtliche Darstellungsweise der Ergebnisse entwickelt. Diese berücksichtigt, dass in Pathologiebefunden entweder nur wenige oder auch eine Vielzahl unterschiedlicher Biomarker erwähnt sein können.

Es wurde sichergestellt, dass die Biomarker-Extraktion insbesondere für die organspezifischen Zusatzmodule *Mamma* und *Kolorektales Karzinom* erfolgt und somit Klassifikatoren für häufige Tumorentitäten verfügbar sind. Die Komponente ist vollständig in die bestehende Pipeline für Pathologiebefunde integriert und somit auch kombiniert mit der Extraktionslogik aus dem HAP2.

5.4 Arbeitspaket 4: Ergebnisvernetzung

AF1 - Averbis: Die von Averbis entwickelten Verfahren und Erkenntnisse wurden erfolgreich in bestehende Softwarearchitekturen mehrerer Landeskrebsregister eingebunden. So kommen die im Rahmen des Projektes entwickelten Extraktionskomponenten der Health Discovery bereits in Baden-Württemberg, Hessen, Rheinland-Pfalz und Nordrhein-Westfalen sowie zwischenzeitlich in Niedersachsen zum Einsatz. Außerdem stehen sie dem Klinischen Krebsregister der Universitätsklinik Freiburg (CCCF) zur Verfügung.

Des Weiteren wurden die erarbeiteten Ergebnisse und Verfahren in mehreren Austauschformaten mit den direkt beteiligten Krebsregistern sowie mit den Landeskrebsregistern Niedersachsen und Nordrhein-Westfalen diskutiert. Zusätzlich wurden die potenziellen Anwendungsbereiche der entwickelten Methoden in einer gesonderten Präsentation dem Krebsregister Bayern vorgestellt und hinsichtlich ihrer Übertragbarkeit erörtert.

Die Ergebnisse fanden zudem Eingang in wissenschaftliche Beiträge, darunter als inhaltlicher Bestandteil in Vorträgen auf der European Network of Cancer Registries (ENCR) – International Association of Cancer Registries (IACR)-Konferenz 2023 sowie der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie (GMDS)-Jahrestagung 2025, was die nationale und internationale Anerkennung der Projektergebnisse fördert.

AF2 – Fraunhofer FKIE: Die Verfahren und Software, die innerhalb des Projektes entwickelt wurden, stehen zur Nutzung durch alle Krebsregister zur Verfügung. Die Implementierung der Algorithmen erfolgte unter Nutzung von Open-Source-Frameworks und als unabhängige Microservices mit definierten standardisierten Schnittstellen, sodass diese in den jeweiligen heterogenen Systemumgebungen der unterschiedlichen Krebsregister eingebunden werden können. Zur Vorbereitung der technischen Übergabe sowie Einbindung in die Informationstechnologie (IT) der Register wurden Workshops mit dem IT- bzw. Stabspersonal des Krebsregisters Baden-Württemberg durchgeführt, um zukünftige operative Input-/Output-Schemata, Datenaustauschformate sowie Anforderungen an die Laufzeit- und Betriebsumgebung festzulegen. Die Übergabe und Einbindung der finalen FKIE-Software-Anteile ist für das erste Quartal 2026 vorgesehen. Eine Weitergabe an andere Krebsregister wird nach Projektabschluss über das Krebsregister Baden-Württemberg koordiniert.

Des Weiteren wurde das algorithmische Funktionsprinzip der Verfahren in AF2 auf mittlerer Abstraktionsebene speziell für das Personal der Krebsregister dokumentiert mit dem Ziel, den Lösungsansatz auch für nicht-technisches Personal, d.h. für die Fachleute für die jeweilige

Domäne und die Dokumentierenden, transparent und verständlich zu machen. Mithilfe der Dokumentation sowie mehreren Workshops wurde das Krebsregister Baden-Württemberg dazu befähigt, bei Bedarf eigenständig moderate Anpassungen bei sich ändernden Randbedingungen (z.B. neue Schnittstellen, Dokumentationsvorgaben, zusätzliche zu untersuchende Gene) vornehmen zu können, ohne dass tiefe Expertise zu den KI-Verfahren erforderlich ist. Für eventuelle Rückfragen steht Fraunhofer FKIE auch nach Projektabschluss zur Verfügung.

6. Gender Mainstreaming Aspekte

6.1 Organisatorische und strukturelle Gender Mainstreaming Aspekte

Genderaspekte wurden organisatorisch im gesamten Projekt durchgängig berücksichtigt, sowohl durch die Vorgabe von Stellenausschreibungen, die dem Allgemeinen Gleichbehandlungsgesetz (AGG) entsprechen, als auch durch das Bemühen um divers-offene Team-Zusammensetzungen.

6.2 Inhaltliche Gender Mainstreaming Aspekte

Die Verbesserung der Datenqualität ist die Voraussetzung für die angemessene Analyse genderspezifischer Aspekte der onkologischen Gesundheitsversorgung, insbesondere der molekularbiologischen Analysen.

7. Partizipation

Der Aspekt der Partizipation war für das Vorhaben nur eingeschränkt relevant. Aufgrund der hohen fachlichen und technischen Komplexität stand zunächst die methodische Entwicklung im Vordergrund, sodass eine systematische Beteiligung weiterer betroffener Zielgruppen nicht umgesetzt wurde. Die Einbindung beschränkte sich auf den fachlichen Austausch mit den unmittelbar beteiligten Expertinnen und Experten aus den Krebsregistern. Eine weitergehende, strukturierte Beteiligung soll jedoch in einem nächsten Schritt bundesweit koordiniert nachgeholt werden, um Erkenntnisse für die langfristige Strategie- und Weiterentwicklung der klinischen Krebsregistrierung zu gewinnen.

8. Diskussion der Ergebnisse, Gesamtbeurteilung

8.1 Arbeitspaket 2 (AF1)

Das übergeordnete Ziel dieses Arbeitspakets war die Entwicklung eines Verfahrens zur möglichst vollständigen und zuverlässigen Extraktion befundrelevanter Merkmale aus freitextlichen Pathologiebefunden. Im Zentrum standen dabei die automatische Erkennung der im onkologischen Basisdatensatz (oBDS) definierten Standardentitäten sowie der organspezifischen Zusatzmodule, die Detektion potenzieller Inkonsistenzen zwischen Freitext und Kodierung sowie die Ermöglichung einer automatisierten Weiterverarbeitung der extrahierten Informationen.

Alle gesetzten Hauptziele konnten im Projektverlauf erreicht werden:

- Eine automatisierte Kodierung ist nun für nahezu alle Tumorentitäten möglich.
- Die Extraktion der relevanten oBDS-Merkmale erfolgt in hoher Qualität und deckt sowohl Standard- als auch Zusatzmodule ab.
- Die Analyse zeigte, dass Inkonsistenzen zwischen Textinhalt und kodierter Information detektiert und systematisch adressiert werden können - sowohl auf technischer Ebene (Modellanpassung) als auch organisatorisch (z. B. durch Rückmeldung an die Register oder Anpassung von Leitlinien).
- Insgesamt konnte gezeigt werden, dass die automatisierte Auswertung von Pathologiebefunden im Rahmen des Projekts praxisreif umgesetzt wurde.

Die automatische Kodierung von Befunden des kolorektalen Karzinoms ist aktuell noch nicht vollständig möglich, da noch nicht alle Felder des organspezifischen Zusatzmoduls extrahiert werden können.

Trotz dieser sehr positiven Ergebnisse besteht in einzelnen Bereichen weiteres Verbesserungspotenzial, insbesondere im Hinblick auf Modellgenauigkeit und Datenbasis:

- Für die Tumorgröße bei DCIS wäre eine Erweiterung des Trainingsdatensatzes hilfreich, um die Unterscheidung zwischen invasiven und nicht-invasiven Karzinomen noch robuster zu gestalten.
- Die Annotation der Hormonrezeptoren im Mamma-Modul war zum Teil unvollständig. Eine erweiterte, konsistente Datenbasis könnte hier zu noch besseren Ergebnissen führen.
- Bei der Extraktion von Stanzenanzahl und -befall ist ein einheitlicher Goldstandard notwendig. Erste Schritte zur Vereinheitlichung der Kodierrichtlinien wurden im Projekt angestoßen und sollten weiterverfolgt werden.
- Multitask-CNNs bleiben trotz initial nur geringer Vorteile ein vielversprechender Ansatz, der mit vollständigeren Annotationen und optimierten Architekturen künftig eine effizientere gemeinsame Modellierung mehrerer Standardentitäten ermöglichen könnte. Auch der Einsatz moderner Verfahren wie Large Language Models (LLMs) bietet Potenzial für zukünftige Optimierungen, etwa durch bessere Kontextverarbeitung oder Zero-/Few-Shot-Lernen bei seltenen Merkmalen.

Insgesamt wird eindrucksvoll gezeigt, dass eine automatische, robuste und registerübergreifende Extraktion zentraler Tumormerkmale aus Freitexten technisch umsetzbar ist und sich in der Praxis erfolgreich einsetzen lässt. Die entwickelte Lösung bildet damit eine tragfähige Grundlage für die weitere Automatisierung in der Dokumentation, sowohl zur Entlastung der Register als auch zur Verbesserung der Datenqualität.

8.2 Arbeitspaket 3 (AF2)

Ziel des Arbeitspakets war die systematische Erschließung bislang nicht strukturierter molekularpathologischer Informationen aus den vorliegenden Befundtexten, um genetisch relevante Informationen, Hauptmutationen und diagnostisch bedeutsame Modifikationen automatisiert identifizieren und bereitstellen zu können.

Alle gesetzten Ziele konnten im Projektverlauf erreicht werden:

- AF2.1: Erkennung von Biomarkern (AP 3.1), so dass auch neu zu dokumentierende Biomarker jederzeit bearbeitet werden können (AP 3.4).
- AF2.2: Extraktion unterschiedlicher spezifischer Patterns, u.a. cDNA- und Protein-Varianten und Exone (AP 3.3), so dass auch neu auftretende Varianten detektiert werden können (AP 3.4).
- AF2.3: Erkennung von zu dokumentierenden Aussagen zu Genen (Zwischenziel von AP 3.2 und 3.4)
- AF2.4: Klassifikation der Ausprägung möglicher Gen-Varianten (AP 3.2) unter Verwendung von generischen Ansätzen, so dass auch neu auftretende Ausprägungen damit erkannt werden können (AP 3.4)
- Als übergreifendes Ziel: Die Entscheidungen der Text-Mining-Verfahren sind für Dokumentierende nachvollziehbar, indem die ursächlichen Textstellen und ausschlaggebenden Wörter auf einen Blick erkennbar sind.

Für eine operative Nutzung sollten in Bezug auf AF2.4 / AP 3.2 noch die folgenden Aspekte finalisiert werden, wie in Kapitel 5, AP 3.1 – 3.4, Schritt 4 beschrieben wurde:

- Bei Vorliegen von problematischer Negation sollte eine Verfeinerung des grammatikalischen Parsings erfolgen, um das Bezugswort der Negation eindeutig feststellen und diese korrekt anwenden zu können. Hierdurch kann der aktuelle Fehler von 0,5% weiter reduziert werden.
- Bei Vorliegen von „vermischten“ Aussagen, die gleichzeitig mehrere Gene oder Varianten beschreiben, sollte die in Kapitel 5, AP 3.1 – 3.4, Schritt 4, Problem 3 definierte Fallunterscheidung angewendet werden.
- Einzelklassifikationen müssen auf der Ebene der Varianten und nicht auf der Ebene der (standardisierten) Gen-Namen zusammengeführt werden, um zwischen unterschiedlichen Ergebnissen (Mutation / Wildtyp) bei mehreren untersuchten Exonen oder Varianten-Arten differenzieren zu können.

Im Ergebnis konnten molekularpathologische Informationen erfolgreich erschlossen und in strukturierter Form nutzbar gemacht werden. Hierzu zählt insbesondere die standardisierte Nennung und aktuelle Beurteilung relevanter Biomarker. Für 90 ausgewählte Biomarker (AP 3.1 – 3.4) bzw. 20 Biomarker (AP 3.5) wurden Untersuchungsergebnisse automatisiert erkannt und extrahiert.

Das daraus gewonnene diagnoserelevante Wissen wird in einer standardisierten Struktur bereitgestellt, was eine einfache Erweiterung um zusätzliche Biomarker ermöglicht. Zudem wurde durch die entwickelten Verfahren die Grundlage für eine nachträgliche Kodierung älterer, bislang unstrukturierter Befunddaten geschaffen.

Diskussion zur Wahl des algorithmischen Text-Mining-Ansatzes in AP3 (AF2)

Da zu Beginn des TeMeK-Vorhabens im AF2 keine umfangreichen annotierten Trainingsdaten verfügbar waren, mithilfe derer Supervised-Machine-Learning-Ansätze (ML) systematisch hätten

trainiert werden können, wurde durch Fraunhofer FKIE ein regel- und schlüsselwort-basiertes Verfahren gewählt und die dazu benötigten Wortschätze sowie Regeln sukzessive aufgebaut und verfeinert. Die entstandenen Klassifikatoren sind den klassischen Natural-Language-Processing-Verfahren zuzuordnen, algorithmisch transparent, sehr schnell und für Menschen gut nachvollziehbar („explainable“).

Stehen zukünftig ausreichend annotierte Trainingsdaten zur Verfügung, wäre es theoretisch möglich, einzelne Schritte der Text-Mining-Prozesskette (siehe Kapitel 5, AP3) durch komplexere KI-Ansätze abzulösen. Die ersten drei der vorgestellten Prozessschritte, nämlich (1) die Erkennung und Standardisierung von Gen-Bezeichnungen, (2) die Extraktion von Pattern für Protein-/cDNA-Varianten, Referenzsequenz-Nummern und Exone sowie (3) die Erkennung von tabellarischen Daten im Rahmen der Relevanz-Klassifikation sollten aber auch weiterhin mit zeichenbasierten Patterns gelöst werden, da diese fehlerfrei und sehr effizient bzgl. Zeit und Hardware-Kosten sind, d.h. ohne spezifische KI-Beschleuniger auskommen.

Potenziell durch komplexere KI-Verfahren ersetzbar wären die Klassifikation der Relevanz (Prozessschritt 3) sowie der Ausprägung (Prozessschritt 4) einer Gen-Nennung. Hier bestünde ein alternatives state-of-the-Art Verfahren darin, mithilfe von kompakten transformerbasierten Embedding-Modellen (nicht-generativen Large-Language-Modellen (LLM)) zunächst geeignete Vektor-Repräsentationen für die Gen-Nennungen zu erzeugen und diese dann mittels eines Supervised-Machine-Learning (ML)-Verfahrens binär zu klassifizieren. Da die Relevanz-Klassifikation (Schritt 3) aber auch klassisch sehr gut funktioniert, bestünde hier kein Mehrwert. Im Gegenteil, der aktuell transparente schlüsselwortbasierte Ansatz würde eher undurchsichtig und könnte bei Aufkommen völlig neuer Begrifflichkeiten nicht einfach durch Erweiterung des Wortschatzes ergänzt werden, sondern müsste neu trainiert werden.

Für die Ausprägungs-Klassifikation (Schritt 4) bestünde ein Vorteil von LLM-basierten Verfahren darin, dass LLMs von Haus aus besser mit Negation umgehen können, da Sprachmodelle ein inhärentes, kontextbezogenes Sprachverständnis besitzen. Allerdings zeigt die Evaluation auf dem Testkorpus mit 21.000 Meldungen, dass das Negations-Problem auch durch den klassischen NLP-Ansatz bereits gut gelöst ist. Da problematische Negation zudem als solche erkannt werden kann, ist es darüber hinaus möglich, das grammatikalische Parsing (Parse-Tree) für solche Sätze gezielt detaillierter durchzuführen, um Fehler zu reduzieren. Somit besteht bzgl. der Negation unserer Einschätzung nach kein Mehrwert in der Nutzung eines LLM-/ML-Ansatzes, der zudem intransparent ist und bzgl. Hw- und Laufzeit-Kosten höher liegt.

Für die Klassifikation der Gesamtausprägung (ebenfalls Schritt 4) gilt außerdem, dass auch ML-basierte Verfahren mit denselben Herausforderung zu kämpfen haben, nämlich der Vermischung von Aussagen zu mehreren Gen-Varianten und unterschiedlichen Genen im selben Hauptsatz (vgl. Kapitel 5, AP 3.1-3.4, Problem Nr. 3, Fall Nr. 3). Eine ML-basierte binäre Klassifikation kann ebenfalls nicht ohne Weiteres verschiedene Varianten- und Gen-Arten im selben Satz korrekt zuordnen. Im Gegensatz dazu bestünde eine echte Alternative in der Verwendung eines sehr großen, generativen LLMs, welches hierfür mit einem ausgefeilten Prompt sowie viel Kontextwissen versorgt werden müsste. Bei diesem Ansatz wird das generative LLM über den Prompt mit präzisen, formalisierten, fachlichen Dokumentationsregeln angeleitet, die Dokumentationstätigkeit unmittelbar nachzuahmen und z.B. direkt alle Varianten im selben Satz auf einmal zu erkennen und bzgl. Ausprägung zu differenzieren. Da generative LLMs aktuell einer rasanten Entwicklung und Verbesserung unterliegen, ist dies in naher Zukunft vorstellbar. Beispielsweise gibt es seit Mitte 2025 einige offene, d.h. Lizenzkostenfreie LLMs (u.a. das Qwen-3-Modell), die in LLM-Benchmarks die benötigten Fähigkeiten demonstriert haben und gleichzeitig

lokal on-premise betrieben werden können, so dass keine personenbezogenen Daten abfließen. Allerdings benötigen diese generativen LLMs aufgrund ihrer Modellgröße spezielle Hardware mit signifikanten Kosten (z.B. Grafikprozessor (Graphics Processing Unit, GPU) NVIDIA H100 mit 80 Gigabyte (GB) Grafikspeicher ca. 35.000 - 40.000 EUR). Obwohl wir in generativer KI grundsätzlich das Potenzial sehen, die gewünschten Informationen zu extrahieren, ist die Definition entsprechender Prompts nicht trivial und eine weitere Verbesserung keineswegs sicher. Die für den klassischen NLP-Ansatz bereits erarbeiteten formalen Regeln sowie Fallunterscheidungen bilden zur Definition der benötigten Prompts auf jeden Fall die optimale Grundlage. Schließlich würde ein generative-KI-basierter Ansatz die im TeMeK-Vorhaben entwickelte regel- und zeichen-basierte Vorverarbeitung und Extraktionen in den Schritten 1 – 3 (von AF2) nicht obsolet machen, sondern nur Unterfälle der hinteren Klassifikationsschritte 3 und 4 ergänzen. Insgesamt sind die Vor- und Nachteile bzgl. der Hardware-Kosten, der Rechenzeit, des Pflegeaufwands bei Neuerungen, der Nachvollziehbarkeit der Vorgehensweise und des Klassifikationsergebnisses sowie der Korrektheit der Ausgabe gut abzuwägen.

Ein deutlich transparenterer, effektiverer und nachhaltigerer Weg wäre es, standardisierend auf die Art und Weise einzuwirken, wie genetische Varianten in molekularpathologischen Befunden berichtet werden, wenngleich dies nicht kurzfristig zu erreichen ist.

9. Verbreitung der Projektergebnisse, Öffentlichkeitsarbeit

- 2023-2026: Das Krebsregister Baden-Württemberg hat für das Projekt TeMeK einen Internetauftritt eingerichtet (https://temek.de)
- 2023/02: FKIE hat eine Pressemeldung zum Weltkrebstag 2023 veröffentlicht, die in medizinischen Fachmedien aufgegriffen wurde. (https://www.fkie.fraunhofer.de/de/newsroom/Presseinformationen/weltkrebstag-2023-temek.html)
- 2023/04: FKIE hat einen Artikel zu TeMeK im Fraunhofer-Magazin veröffentlicht, welcher frei zugänglich ist: Fraunhofer-Magazin 4.2023 (S. 77): "Künstliche Intelligenz für eine bessere Medizin" – Unterabschnitt „KI für die Krebsregister“. (https://www.fraunhofer.de/s/ePaper/Magazin/2023/04/index.html#78)
- 2023/11: Halber M: Textmining of pathology reports for uniform classifications (TeMeK). Vortrag, Workshop "IT Tools and novel AI approaches for cancer registration", ENCR-IACR 2023 Scientific Conference, Granada, Spanien, 13.11.2023.
- 2023/11: Halber M: TextMining von Meldungstexten für einheitliche Klassifikationen (TeMeK). Vortrag, Fest-Symposium "Gemeinsam zum Erfolg – Ein Datenschatz für die onkologische Forschung und Versorgung in Deutschland" der ADT, der Deutschen Krebsgesellschaft (DKG), der Deutschen Krebshilfe (DKH), des Deutschen Netzwerks Versorgungsforschung (DNVF), der Plattform 65c, der GEKID, des Hauses der Krebs-Selbsthilfe und des Robert Koch-Instituts (RKI), Berlin, 20.11.2023.
- 2025/09: Halber M: Erkenntnisse aus TeMeK: TextMining von Meldungstexten für einheitliche Klassifikationen. Vortrag, Jahrestagung der GMDS, Jena, 09.09.2025.
- 2025/12: FKIE hat einen Artikel zu TEMEK im Jahresbericht 2025 veröffentlicht, welcher frei zugänglich ist: FKIE-Jahresbericht 2025 (S. 15): „Krebsmutationen mithilfe von Data Mining auf die Spur kommen“. (https://www.fkie.fraunhofer.de/de/ueber-fkie.html)
- 2025/12: FKIE hat das algorithmische Funktionsprinzip der in AF2 implementierten Verfahren auf mittlerer Abstraktionsebene speziell für das Personal von Krebsregistern dokumentiert mit dem Ziel, den Lösungsansatz auch für nicht-technisches Personal, d.h. für die Fachleute für die jeweilige Domäne und die Dokumentierenden, transparent und verständlich zu machen. Die Dokumentation wird gemeinsam mit der Software verteilt und ist für alle Krebsregister verfügbar.
- 2026/02: FKIE hat eine Pressemeldung zum Weltkrebstag 2026 veröffentlicht und plant im Anschluss weitere Social-Media-Aktivitäten (u.a. via LinkedIn).

10. Verwertung der Projektergebnisse (Fortführung, Nachhaltigkeit, Transferpotenzial)

AF1 - Averbis: Die von Averbis entwickelten Verfahren und Erkenntnisse wurden erfolgreich in bestehende Softwarearchitekturen mehrerer Landeskrebsregister eingebunden. So kommen die im Rahmen des Projektes entwickelten Extraktionskomponenten der Health Discovery bereits in Baden-Württemberg, Hessen, Rheinland-Pfalz und Nordrhein-Westfalen sowie zwischenzeitlich in Niedersachsen zum Einsatz. Außerdem stehen sie dem Klinischen Krebsregister der Universitätsklinik Freiburg (CCCF) zur Verfügung.

Des Weiteren wurden die erarbeiteten Ergebnisse und Verfahren in mehreren Austauschformaten mit den direkt beteiligten Krebsregistern sowie mit den Landeskrebsregistern Niedersachsen und Nordrhein-Westfalen diskutiert. Zusätzlich wurden die potenziellen Anwendungsbereiche der entwickelten Methoden in einer gesonderten Präsentation dem Krebsregister Bayern vorgestellt und hinsichtlich ihrer Übertragbarkeit erörtert.

Die Ergebnisse fanden zudem Eingang in wissenschaftliche Beiträge, darunter als inhaltlicher Bestandteil in Vorträgen auf der ENCR-IACR-Konferenz 2023 sowie der GMDS-Jahrestagung 2025, was die nationale und internationale Anerkennung der Projektergebnisse fördert.

AF2 – Fraunhofer FKIE: Die Verfahren und Software, die innerhalb des Projektes entwickelt wurden, stehen zur Nutzung durch alle Krebsregister zur Verfügung. Die Implementierung der Algorithmen erfolgte unter Nutzung von Open-Source-Frameworks und als unabhängige Microservices mit definierten standardisierten Schnittstellen, sodass diese in den jeweiligen heterogenen Systemumgebungen der unterschiedlichen Krebsregister eingebunden werden können. Zur Vorbereitung der technischen Übergabe sowie Einbindung in die Register-IT wurden Workshops mit dem IT- bzw. Stabspersonal des Krebsregisters Baden-Württemberg durchgeführt, um zukünftige operative Input-/Output-Schemata, Datenaustauschformate sowie Anforderungen an die Laufzeit- und Betriebsumgebung festzulegen. Die Übergabe und Einbindung der finalen FKIE-Software-Anteile ist für das erste Quartal 2026 vorgesehen. Eine Weitergabe an andere Krebsregister wird nach Projektabschluss über das Krebsregister Baden-Württemberg koordiniert.

Bzgl. der inhaltlichen KI-/Textmining-Verfahren kann festgehalten werden, dass die Auswertung und Klassifikation molekularpathologischer Befunde bzgl. genetischer Varianten (AF2) mittels Text-Mining grundsätzlich sehr gut möglich ist und zudem so generisch umgesetzt werden kann, dass sie auch bei neu aufkommenden Biomarkern und Varianten nutzbar bleibt (siehe AP 3.4). Aufgrund der Neuheit der genetischen Untersuchungsmethoden ist die Dokumentationsweise für molekulargenetische Befundergebnisse durch die klinischen Krebsregister jedoch noch nicht vollständig gefestigt. Es kommen immer wieder neue Anforderungen an die Dokumentation hinzu, z.B. werden weitere Informations- und Mutationsarten relevant, die aus den Befunden als strukturierte Zusatzinformation herausgezogen werden sollen. Gleichzeitig treten durch die Auswertung großer Befundmengen Sonderfälle und neue weitergehende Fragestellungen auf, die bzgl. einer einheitlichen Dokumentation zwischen den Fachexpertinnen und Fachexperten der klinischen Krebsregister zunächst abzustimmen sind. Aus diesem Grund ist die im Projekt entwickelte Toolunterstützung nicht als final anzusehen, sondern unterliegt einer ständigen Weiterentwicklung parallel zur Ausdifferenzierung / Anpassung der Dokumentationsrichtlinien. Um bei erforderlichen Anpassungen ggf. unterstützen zu können, soll das beim Fraunhofer-Instituts für Kommunikation, Informationsverarbeitung und Ergonomie FKIE aufgebaute Domänenwissen erhalten und noch weiter ausgebaut werden. Zudem wird eine langfristige Kooperation mit den Krebsregistern angestrebt, auf deren Basis auch zukünftig Workshops durchgeführt, wissenschaftliche Fragen zu Textmining und neuen generative KI-Verfahren erörtert und

Erkenntnisse ausgetauscht werden können. Entsprechende Vereinbarungen und Data Sharing Agreements wurden bereits ausgearbeitet.