Aus dem Department für Nutztiere und öffentliches Gesundheitswesen
der Veterinärmedizinischen Universität Wien

Zentrum für Tierernährung und Tierschutzwissenschaften (Leitung: Univ.-Prof. Dr.sc.agr. Qendrim Zebeli)

Vergleich manueller und bildbasierter Erhebungen von Integumentschäden bei Mastschweinen nach der Schlachtung

Diplomarbeit

Veterinärmedizinische Universität Wien

vorgelegt von

Theresa Heindl

Wien, im Oktober 2025

BetreuerInnen:

Ass.-Prof. Dr. vet. med. Johannes Baumgartner
Zentrum für Tierernährung und Tierschutzwissenschaften
Department für Nutztiere und öffentliches Gesundheitswesen der Veterinärmedizinischen Universität Wien

Dr. vet. med. Kristina Maschat
Zentrum für Tierernährung und Tierschutzwissenschaften
Department für Nutztiere und öffentliches Gesundheitswesen der Veterinärmedizinischen Universität Wien

Dr. Christina Pfeiffer
Zentrum für Tierernährung und Tierschutzwissenschaften
Department für Nutztiere und öffentliches Gesundheitswesen der Veterinärmedizinischen Universität Wien

Begutachter:

Ao.Univ.-Prof. Dr.med.vet. Peter Paulsen, Dipl. ECVPH

Eigenständigkeitserklärung

Hiermit erkläre ich, dass ich die vorgelegte Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Alle übernommenen Textstellen aus fremden Quellen wurden kenntlich gemacht.

Ich habe die entscheidenden Arbeiten selbst durchgeführt und alle zu erarbeitenden Tätigkeiten mit ihrem Beitrag zur Arbeit angeführt.

Die vorliegende Arbeit wurde nicht an anderer Stelle eingereicht oder veröffentlicht.

Wien, den [14.10.2025] Theresa Heindl

ZUSAMMENFASSUNG

In der vorliegenden Arbeit wird ein Vergleich zwischen manuellen und bildbasierten Erhebungen von Integumentschäden bei Mastschweinen nach der Schlachtung durchgeführt. Das Ziel ist es, die Übereinstimmung beider Methoden zu bewerten und das Potenzial einer automatisierten Bildanalyse für das Tierwohlmonitoring am Schlachthof zu ermitteln. Untersucht werden dabei insgesamt 765 Schweine aus verschiedenen Herkunftsbetrieben am Schlachthof Higelsberger in Schwertberg, OÖ. Für jedes Schwein erfolgt eine manuelle Dokumentation durch geschulte BeobachterInnen, sowie eine Annotierung der verschiedenen Hautläsionen mittels bildbasierter Technik.

Bewertet werden dabei unter anderem die Kategorie, Größe und Tierschutzrelevanz der Läsionen, sowie eine Ermittlung der absoluten und relativen Häufigkeiten.

Es werden außerdem Beobachterabgleiche auf Basis eines bildbasierten Systems durchgeführt um die Konsistenz und Vergleichbarkeit zwischen zwei BeobachterInnen zu überprüfen und zu bewerten.

Die statistische Auswertung umfasst neben der Berechnung von Cohen's Kappa, auch die Ermittlung der Genauigkeit und Präzision. Die Ergebnisse zeigen je nach Läsionskategorie eine moderate bis hohe Übereinstimmung, während andere Kategorien geringere Werte erreichen.

Zudem ist es im Rahmen des Projektes vorgesehen, eine sogenannte *Ground Truth* innerhalb eines Expertenteams zu entwickeln, um einen möglichst objektiven Referenzstandard für die Basis eines Algorithmus zur automatischen Erkennung von Läsionen zu schaffen.

Die Arbeit kommt zu dem Schluss, dass eine Datenerhebung am Schlachthof auf Basis von KI eine sinnvolle Ergänzung zur manuellen Untersuchung darstellen kann, insbesondere ergänzend zu bereits bestehenden Methoden, jedoch noch Optimierungen nötig sind, um eine gleichbleibend hohe Genauigkeit zu gewährleisten.

ABSTRACT

This study compares manual and image-based assessments of integument damage in fattening pigs after slaughter. The aim is to evaluate the consistency of both methods and to determine the potential of automated image analysis for animal welfare monitoring at slaughterhouses.

A total of 765 pigs from various farms are examined at the Higelsberger slaughterhouse in Schwertberg, Upper Austria. For each pig, manual documentation is carried out by trained observers, and the various skin lesions are annotated using image-based technology.

Among other things, the category, size and animal welfare relevance of the lesions are evaluated, and the absolute and relative frequencies were determined.

Observer comparisons are carried out as well on the basis of an image-based system to check and evaluate the consistency and comparability between two observers.

In addition to calculating Cohen's Kappa, the statistical evaluation also includes determining accuracy and precision. The results show moderate to high agreement depending on the lesion category, while other categories achieve much lower values.

In addition, the project aims to develop a so-called *ground truth* within a team of experts in order to establish a reference standard for the algorithm. This *ground truth* is to be as objective as possible to serve as the base of the algorithm.

The study concludes that AI-based data collection at slaughterhouses can be a useful addition to manual inspection, especially when used in conjunction with existing methods, but that optimisations are still needed to ensure consistently high accuracy.

Inhaltsverzeichnis

1. Einleitung 1
1.1 Zielsetzung und Hypothesen 3
2. Material, Methoden und Tiere 4
2.1 Datenerfassung am Schlachthof Higelsberger 5
2.2 Kamerasystem der Firma „CLK GmbH“ 7
2.3 Annotierung der Bilder mittels CVAT 8
2.4 Durchführung der Beobachterabgleiche durch zwei beteiligte Beobachterinnen 15
2.5 Statistische Auswertung der Daten 16
2.6 Optimierung der Ground Truth im Expertenteam 18
3. Ergebnisse 20
3.1 Prävalenzen von Hautläsionen erhoben am Schlachthof Higelsberger 20
3.2 Prävalenzen von manuell am Schlachthof erhobenen Hautläsionen im Vergleich mit bildbasierter Bewertung 21
3.3 Beobachterabgleich zwischen zwei beteiligten Beobachterinnen 23
3.3.1 Ergebnisse der Beobachterabgleiche vom 23.07.2024 und 30.07.2024 23
4. Diskussion 28
4.1 Herausforderungen bei der Klassifikation von Hautläsionen als prozessbedingt oder tierschutzrelevant 28
4.2 Optimierungsschritte des Beobachterabgleiches zwischen zwei Bewerterinnen und Auswirkungen auf die berechnete Übereinstimmung 32
4.3 Schwierigkeiten und Optimierungsmöglichkeiten bei der Erstellung der Ground Truth 33
4.4 Schlussfolgerung 35
5. Anhang 36
5.1 Literaturverzeichnis 36
5.2 Abbildungsverzeichnis 38

1. Einleitung

In den letzten Jahrzehnten konnte eine Intensivierung der Schweinehaltung, bei gleichzeitiger Steigerung der Produktionseffizienz beobachtet werden.

Während Tiergesundheit schon länger im Fokus steht, rückt seit einiger Zeit auch das Wohlbefinden der Tiere in den Mittelpunkt.

Es können zwar entlang der gesamten Produktionskette (landwirtschaftlicher Betrieb, Transport und Schlachtung) Tierwohlindikatoren systematisch erfasst werden, die Bewertung am Schlachthof bietet dabei jedoch einige Vorteile. Hier kommen viele Tiere aus verschiedenen Betrieben unter den gleichen Bedingungen zusammen, wodurch eine besser standardisierte Beurteilung ermöglich wird. Ein weiterer Vorteil ist außerdem, dass aus Sicht der Biosicherheit keine Gefahr der Rückübertragung von Krankheiten in die Tierbestände erfolgt.

Aus diesen Gründen werden schon seit längerem nicht nur am lebenden Tier auf Betrieben, sondern auch direkt am Schlachthof Tierwohl-Indikatoren erfasst. Im Rahmen der Schlachtkörper- und Fleischuntersuchung (SFU) werden die Tiere ante und post mortem auf Veränderungen untersucht. Jeder Code besteht aus einem Buchstaben, welcher die Befundgruppe beschreibt und einer dreistelligen Zahl, für den genauen Befund. Beispielsweise steht der Code C130 für hochgradige Abmagerung, ein klarer Tierschutzverstoß.

Auch bereits bestehende Daten aus der SFU spielen eine wichtige Rolle, da sie auch zunehmend im Kontext des Tierwohls interpretiert werden können – etwa im Hinblick auf Hautläsionen, Abszesse und/oder Gelenkveränderungen.

Bislang erfolgt die Beurteilung der Schlachtkörper durch amtliche TierärztInnen und MitarbeiterInnen der „Österreichischen Fleischkontrolle“ (ÖFK). Dabei kann es jedoch zu deutlichen Varianzen in der Bewertung kommen, die auf subjektive Wahrnehmung und individuelle Erfahrung der Beurteilenden zurückzuführen sind [1].

Die automatisierte Erfassung könnte eine bessere standardisierte Bewertung tierbezogener Merkmale ermöglichen. Dass der Einsatz künstlicher Intelligenz (KI) zur Erkennung tierschutzrelevanter Indikatoren realisierbar ist, zeigen erste Studienergebnisse. So können in einer Studie aus dem Jahr 2025 mithilfe eines künstlichen neuronalen Netzwerkes Ohrläsionen bei Schweinen mit hoher Genauigkeit direkt am Schlachtkörper identifiziert werden [2].

Eine bereits 2020 durchgeführte Pilotstudie zeigt, dass ein kamerabasiertes System zur Erkennung von Ohr- und Schwanzverletzungen eine hohe Übereinstimmung mit der herkömmlichen Bildbewertung aufweist [1].

Diese Ergebnisse belegen, dass Sensor-basierte und KI-unterstützte Verfahren tatsächlich das Potential haben, tierschutzrelevante Läsionen zu erfassen und zu verarbeiten.

Im Rahmen des EU Projektes „aWISH“, im Zuge dessen die vorliegende Arbeit verfasst ist, stellt sich demnach die Frage, ob sich mithilfe von moderner KI-basierter Bildverarbeitung Methoden zur Bewertung tierschutzrelevanter Läsionen an Schweinen entwickeln lassen können [3]. Dabei wird auch untersucht, inwiefern der Einsatz von künstlicher Intelligenz valide Rückschlüsse auf das Tierwohl zulässt und wie diese Ergebnisse im Vergleich zur konventionellen Erfassung vor Ort einzuordnen sind.

Der Fokus liegt dabei insbesondere auf der Beurteilung von Hautläsionen. Diese können während der gesamten Lebensdauer der Tiere durch verschiedene Faktoren und Einflüsse wie beispielsweise eine mangelhafte Haltung oder schlechte Umweltbedingungen entstehen und sind somit ein hervorragender Indikator zur Beurteilung der ganzheitlichen Tiergesundheit und des Tierwohls.

Dadurch ist es schlussendlich auch möglich, direktes und konstruktives Feedback an alle Beteiligten der gesamten Lieferkette zu geben und somit nachhaltig das Tierwohl zu verbessern.

1.1 Zielsetzung und Hypothesen

Das übergeordnete Ziel der vorliegenden Arbeit ist es, eine Grundlage für die Entwicklung KI-gestützter Verfahren zur analogen und darauffolgenden automatisierten Erkennung von Hautläsionen an Schweinen zu bilden. Dabei steht die Definition der sogenannten *Ground Truth,* welche als objektive Referenz für die spätere Validierung eines Algorithmus zur automatisierten Bildbewertung dienen soll, im Fokus. Dazu werden Hautläsionen direkt am Schlachthof erhoben, dokumentiert, klassifiziert und anschließend mit den entsprechenden Bilddaten abgeglichen.

Die Erhebung der Prävalenzen verschiedener Hautläsionen direkt am Schlachthof bildet hierbei die erste zentrale Zielstellung.

Ein zweiter Schwerpunkt liegt auf der Beurteilung der entsprechenden am Schlachthof aufgenommenen Bilddaten mittels eines Computer Vision Annotation Tool (CVAT).

Hierbei soll vor allem untersucht werden, inwieweit die Klassifikation der verschiedenen Läsionen durch zwei menschliche Beobachter konsistent ist und übereinstimmt.

Außerdem wird der Ablauf der *Ground Truth*-Entwicklung im Expertenteam dargestellt und damit einhergehende methodische Herausforderungen beschrieben.

Im dritten Schritt erfolgt ein Vergleich zwischen der bildbasierten Bewertung und der konventionellen Erhebung direkt am Schlachthof. Es wird analysiert, welche Läsionen durch bildgestützte Verfahren im Vergleich zur manuellen Befundung am Schlachtband dokumentiert werden können und inwiefern sich die Bewertung der entsprechenden Läsionen zwischen beiden Methoden unterscheidet.

Diese drei Aufgabenstellungen bilden die Grundlage für die im Projekt aWISH angestrebte Algorithmus-basierte Erkennung tierschutzrelevanter Läsionen an Schweinen und ermöglicht eine Einschätzung zur Anwendbarkeit von künstlicher Intelligenz im Tierwohlkontext am Schlachthof.

2. Material, Methoden und Tiere

Das Projekt „aWISH“ steht für „animal welfare indicators at the slaughterhouse“ und beinhaltet sowohl Schweine- als auch Hühnermastbetriebe in der gesamten EU.
Es umfasst dabei 24 Partner aus insgesamt elf EU-Ländern - darunter auch Österreich - und erstreckt sich dabei über die komplette Produktionskette [3].
Die vorliegende Arbeit ist Teil dieses Projektes und beschäftigt sich mit der Erfassung von Hautläsionen bei Schweinen am Schlacht- und Zerlegebetrieb Higelsberger GmbH in Schwertberg, Oberösterreich. Dort werden ausschließlich Schweine geschlachtet.
Die durchschnittliche Schlachtgeschwindigkeit am Betrieb liegt bei ungefähr 144 Tieren pro Stunde, was sich an einem typischen Schlachttag auf eine Gesamtzahl von etwa 800 bis 1000 Schweinehälften summiert (Rudolf Baderer, persönliche Mitteilung, 22.04.2025).
Die am Schlachthof in Schwertberg ankommenden Tiere stammen aus umliegenden Mastbetrieben und werden üblicherweise von Montag bis Freitag zwischen 6:00 und 12:30 Uhr der Schlachtung zugeführt. Die Betäubung erfolgt im Betrieb Higelsberger elektrisch in einer vom Betreiber selbst entwickelten Betäubungsanlage.
Die angelieferten Schweine sind das Produkt einer Dreirassenkreuzung, mit den Mutterrassen Edelschwein, Landrasse und der Terminalrasse Pietrain, wie sie typischerweise in der Schweinemast in Österreich eingesetzt wird.
Neben der konventionellen, am gesetzlichen Mindeststandard orientierten Haltung werden immer dienstags auch sogenannte „Fair zum Tier“ Schweine geschlachtet, welche aus der Haltungsform „Mehr Tierwohl – sehr gut“ stammen. Zur Datenerhebung für diese Arbeit werden Tiere aus beiden Haltungsformen herangezogen.
Die Haltungsform „Mehr Tierwohl – sehr gut“ im Rahmen des AMA-Gütesiegels geht dabei deutlich über die gesetzlichen Mindestanforderungen der konventionellen Schweinehaltung hinaus.
Sie umfasst unter anderem mehr Platz, gentechnikfreie Fütterung sowie Kastration nur unter Vollnarkose und Analgesie [4]. Außerdem ist auch das Kupieren der Schwänze verboten und es werden ausschließlich Schweine mit Ringelschwänzen gehalten [4].

2.1 Datenerfassung am Schlachthof Higelsberger

Für die vorliegende Arbeit werden an sieben Erhebungstagen zwischen Juni und Oktober 2024 insgesamt 765 Mastschweine direkt am Schlachtband beurteilt. Die Erhebungen finden dabei jeweils im Zeitraum zwischen 9:00 und 11:30 Uhr statt. Es sind dabei immer dieselben zwei trainierten Personen an der Datenerfassung beteiligt.
Die Aufnahme und Klassifikation der Daten erfolgt dabei unmittelbar nach dem Abflämmen und sobald die Tiere den sogenannten Weißbereich des Schlachtprozesses erreichen.
Um die Durchführung der Datenaufnahme zu erleichtern, wurde vorab eigens für dieses Projekt ein Arbeitsplatz direkt am Schlachtband errichtet, um einen direkten Blick auf die geschlachteten Schweine zu ermöglichen.
Der Arbeitsplatz besteht aus einem höhenverstellbaren Podest und ermöglicht eine vollständige Sicht auf die dorsale Körperseite der Tiere. Pro Schwein steht eine Beurteilungszeit von rund zehn Sekunden zur Verfügung. Daher liegt der Fokus auf der Dokumentation von jeweils jener Hautläsion, die bei einem Schwein visuell am auffälligsten bzw. am prominentesten ist.
Erfasst werden dabei:

- Schlachtnummer (fortlaufend)
- Größe der Hautläsion (cm), gemessen mit einem geeigneten Maßband
- Kategorie der Hautläsion

Zur Dokumentation und Klassifikation der Läsionen und anschließender Digitalisierung mit Microsoft Excel wird eine standardisierte Tabelle (Tabelle 11) verwendet, welche auch als Grundlage für die statistische Auswertung dient und im Anhang dieser Arbeit einsehbar ist. Zur Auswertung der Prävalenzen der Hautläsionen am Schlachthof wird anschließend die absoluten Häufigkeiten ebenso mittels Microsoft Excel berechnet.
Um eine einheitliche Klassifikation der Hautläsionen zu gewährleisten, wird außerdem vorab definiert, welche Merkmale diese jeweils aufweisen muss. Dies geschieht gemäß der nachfolgenden Tabelle 1, welche die Definition der verschiedenen Hautläsionen beschreibt.

Tabelle 1: Definition der Hautläsionen (zur Verfügung gestellt von Dr. med. vet. Kristina Maschat)

Schlachtnummer:	Auf den Schweinekörper gestempelte aufeinanderfolgende Nummer (letzten vier Ziffern). Sie dient der Identifizierung und Nachverfolgung von Schlachtkörpern und wird auf jeder Hälfte des Schlachtkörpers angebracht.
Bisswunde (n):	Einzelne oder gruppierte (zahnreihenförmige) Wunde mit Anzeichen einer Entzündung.
Alter Kratzer (n):	Oberflächliche alte Wunde, bei der die Haut nicht komplett perforiert ist und nur die Epidermis und Dermis betroffen sind. Anzeichen einer Entzündung sind vorhanden und die Veränderung erscheint bräunlich.
Frischer Kratzer (n):	Oberflächliche frische Wunde, bei der die Haut nicht vollständig perforiert ist, sondern nur Epidermis und obere Dermis betroffen sind. Anzeichen einer Entzündung sind vorhanden und die Veränderung erscheint rötlich/violett.
Alte und tiefe Wunde (n):	Die Haut ist vollständig perforiert und alle Hautschichten sind betroffen. Entzündungsanzeichen sind vorhanden und auch Gewebeverlust kann vorhanden sein. Die veränderten Hautbereiche erscheinen bräunlich.
Frische und tiefe Wunde (n):	Die Haut ist vollständig perforiert und alle Hautschichten sind betroffen. Entzündungsanzeichen sind vorhanden und auch Gewebeverlust kann mit der Verletzung einhergehen. Die veränderten Hautbereiche erscheinen rötlich beziehungsweise violett.
Prellung (n):	Violette oder rote Verfärbung der Haut, welche durch Kapillarblutung verursacht wurde.
Läsion verursacht durch menschlichen Umgang (n):	Alle Arten von Hautläsionen, die durch den menschlichen Umgang verursacht werden, mit Ausnahme der unsachgemäßen Anwendung von Tätowiereisen.
Krankheit (ja/nein):	Alle Arten von Ausschlag, welcher auf eine Krankheit hinweisen könnte.
Ungewiss/Unsicher (n):	Alle Arten von Veränderungen, deren Quelle nicht klar ist (einschließlich Veränderungen, welche durch den Schlachtprozess verursacht wurden).
Größe (xy x xy):	Breite und Länge der Veränderung in Zentimeter gemessen (x x y).

2.2 Kamerasystem der Firma „CLK GmbH“

Für die bildbasierte Erfassung und Analyse der Schlachtkörper kommt das hochauflösende Kamerasystem „PigInspector“ der Firma CLK GmbH zum Einsatz (https://www.clkgmbh.de/food/piginspector).

Zur Messung der Schwanzlängen wird eine Optonic 3D-Stereo-Kamera mit einer Auflösung von 2,3 MPx verwendet. Zusätzlich werden Farbkameras der Firma IDS mit einer Auflösung von 5 MPx integriert, um verschiedene Perspektiven der Tiere zu erfassen (Christoph Bonk, persönliche Mitteilung 24.02.2025).

Die eingesetzten Objektive sind dabei speziell auf die jeweiligen Kamerapositionen abgestimmt. Während für die Kopf-, Schinken- und Seitenkameras eine Brennweite von 12 mm verwendet wird, kommt für die Rückenkamera ein Objektiv mit 8 mm Brennweite zum Einsatz (Christoph Bonk, persönliche Mitteilung, 24.02.2025).

Zur Optimierung der Bildqualität unter den gegebenen Bedingungen wird ein gepulstes Beleuchtungssystem eingesetzt, das von CLK für den Einsatz in Schlachthöfen entwickelt wurde. Obwohl das verwendete Licht für das menschliche Auge wie ein Dauerlicht erscheint, arbeitet es mit hochfrequenten Impulsen, vergleichbar mit einem schnellen Blitz. Dies verbessert die Aufnahmebedingungen, minimiert Bewegungsunschärfe und optimiert die Kontraste, um eine bessere Analyse zu ermöglichen.

Die maximale Leistungsaufnahme des Beleuchtungssystems beträgt 100 Watt (Christoph Bonk, persönliche Mitteilung, 24.02.2025).

Die Kameras wurden dabei so installiert, dass sie Bilder der Schweine nach dem Ausbluten, dem Brühen, der Borstenentfernung und dem Flämmen aufgenommen werden. Das ist zur nachfolgenden Beurteilung der Schlachtkörper demnach wichtig, da somit grobe Verschmutzungen entfernt und mögliche tierschutzrelevante Läsionen besser gesehen und dokumentiert werden können.

Zur anschließenden Bildbewertung im Rahmen dieser Arbeit werden der Einfachheit halber die Bilder herangezogen, welche die Rückenansicht der Tiere zeigen.



Abbildung 1: Kamerasystem „PigInspector“ der Firma CLK GmbH am Standort Schwertberg

2.3 Annotierung der Bilder mittels CVAT

Zur Annotierung der am Schlachthof mittels des Kamerasystems „PigInspector“ aufgenommenen Bilder von Schlachtschweinen kommt das sogenannte Computer Vision Annotation Tool (CVAT) zum Einsatz [5]. Dabei handelt es sich um ein Programm zur Bildannotation. Die Bilder werden dabei von der CLK GmbH (CLK) zur Verfügung gestellt. Insgesamt werden sieben ‚Tasks‘, mit jeweils 80 bis 120 Bildern annotiert. Die Aufnahmen zeigen dabei jeweils den Rumpf der Tiere in Dorsalansicht. Abbildung 2 zeigt dabei eine Beispielaufnahme eines Schweineschlachtkörpers aus dem CVAT System.

Bei dem Prozess der Annotierung wird jede im Bild identifizierte Läsion von einer geschulten Person lokalisiert, durch ein Rechteck eingegrenzt und anschließend einer vordefinierten Kategorie (Label) zugeordnet. Die Labels orientieren sich an einem zuvor festgelegten Klassifikationsschema, welches auf dem manuellen Erfassungsprotokoll für Haut- und Schwanzläsionen basiert.

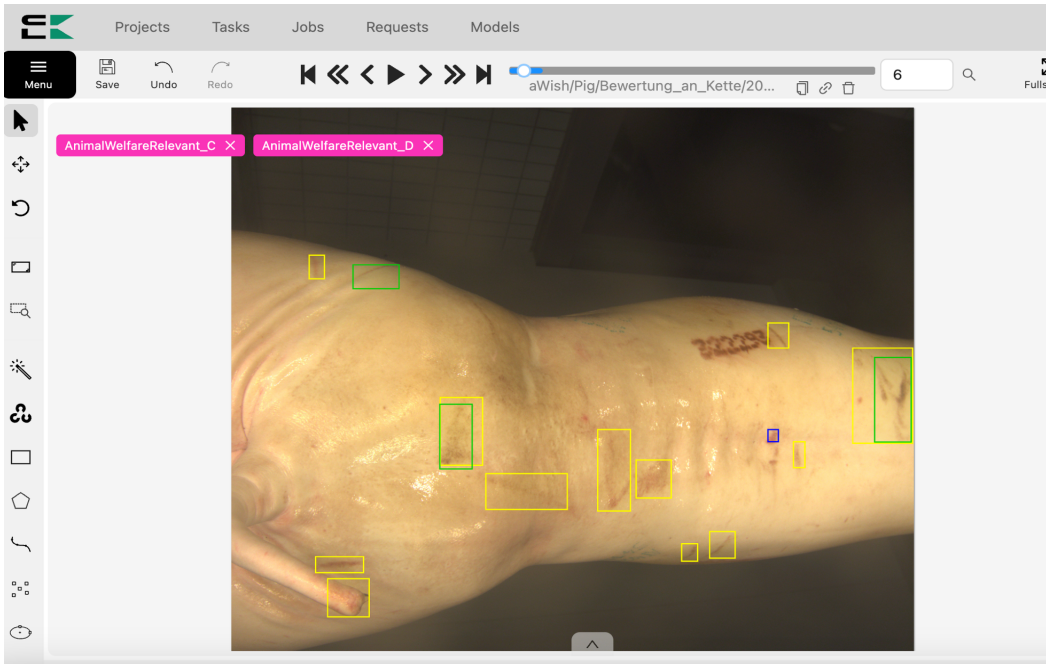


Abbildung 2: dorsale Ansicht eines Schweineschlachtkörpers im CVAT System mit bereits annotierten Läsionen

Die Annotation erfolgte durch jene zwei Beobachterinnen, welche auch an der manuellen Datenaufnahme am Schlachthof beteiligt waren. Insgesamt wurden auf diese Weise 1069 Schlachtschweine beurteilt.

In den Tabellen 2 und 3 werden die verwendeten Labels für die Haut- und Schwanzläsionen und ihre jeweilige Definition dargestellt. Außerdem finden sich unterhalb der jeweiligen Tabelle Beispielbilder zu den häufigsten Veränderungen markiert mit einem Rechteck aus dem CVAT System. Abbildungen 3 bis 8 zeigen dabei typische Hautläsionen. Die Abbildungen 9 bis 14 stellen dagegen typische Schwanzläsionen dar.

Tabelle 2*:* Labels und Definition der Hautläsionen zur Bildannotierung in CVAT

Bezeichnung des Labels	**Definition**
SkinBiteMark	Einzelne oder gruppierte (zahnreihenförmige) Wunde mit Anzeichen einer Entzündung.
SkinOldScratch	Oberflächliche alte Wunde, bei der die Haut nicht komplett perforiert ist und nur die Epidermis und Dermis betroffen sind. Anzeichen einer Entzündung sind vorhanden und die Veränderung erscheint bräunlich.

SkinFreshScratch	Oberflächliche frische Wunde, bei der die Haut nicht vollständig perforiert ist, sondern nur Epidermis und obere Dermis betroffen sind. Anzeichen einer Entzündung sind vorhanden und die Veränderung erscheint rötlich/violett.
SkinOldDeepWound	Die Haut ist vollständig perforiert und alle Hautschichten sind betroffen. Entzündungsanzeichen sind vorhanden und auch Gewebeverlust kann vorhanden sein. Die veränderten Hautbereiche erscheinen bräunlich.
SkinFreshDeepWound	Die Haut ist vollständig perforiert und alle Hautschichten sind betroffen. Entzündungsanzeichen sind vorhanden und auch Gewebeverlust kann mit der Verletzung einhergehen. Die veränderten Hautbereiche erscheinen rötlich beziehungsweise violett.
SkinBruise	Violette oder rote Verfärbung der Haut, welche durch Kapillarblutung verursacht wurde.
SkinHumanHandling	Alle Arten von Hautläsionen, die durch den menschlichen Umgang verursacht werden, mit Ausnahme der unsachgemäßen Anwendung von Tätowiereisen.
SkinDisease	Alle Arten von Ausschlag, welcher auf eine Krankheit hinweisen könnte.
SkinRash	Alle Arten von Ausschlag, welcher auf eine Krankheit hinweisen könnte.
SkinUncertain	Alle Arten von Veränderungen, deren Quelle nicht klar ist (einschließlich Veränderungen, welche durch den Schlachtprozess verursacht wurden).
SkinProcessedCaused	Durch Verarbeitung verursachte Läsion

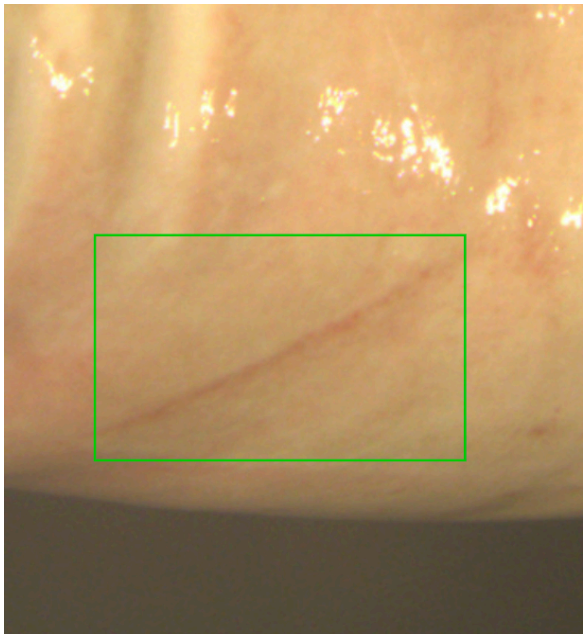


Abbildung 3: Skin Fresh Scratch

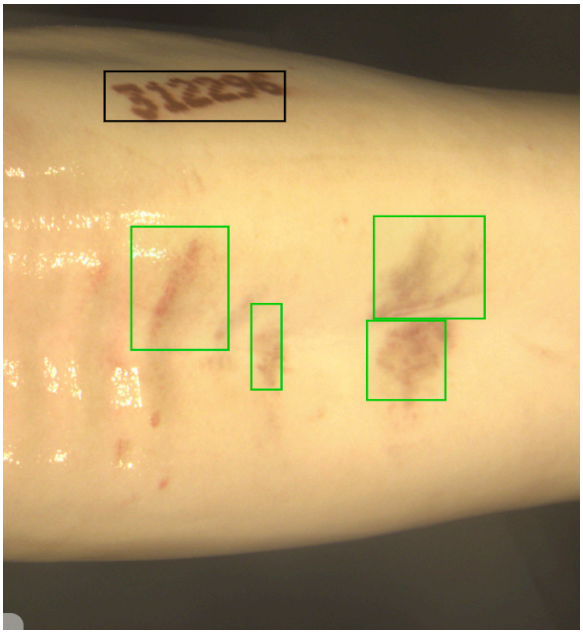


Abbildung 4: Skin Bruise

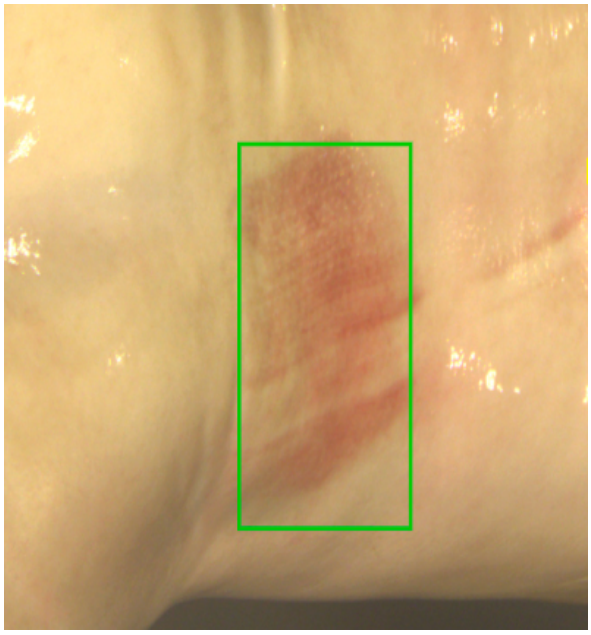


Abbildung 5: Skin Process Caused

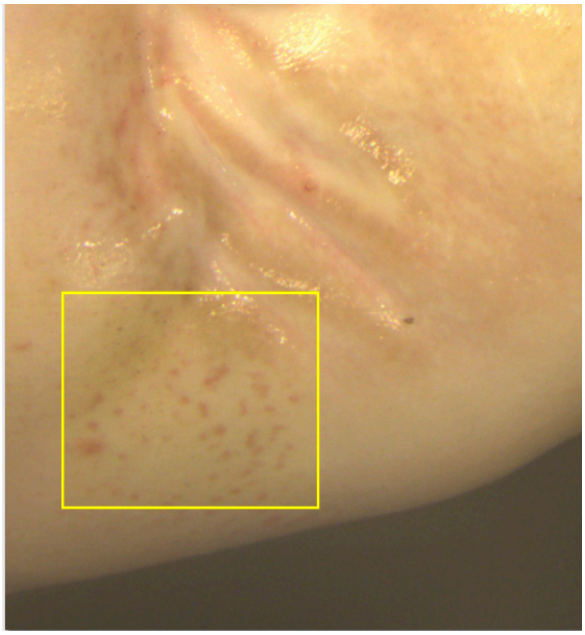


Abbildung 6: Skin Rash

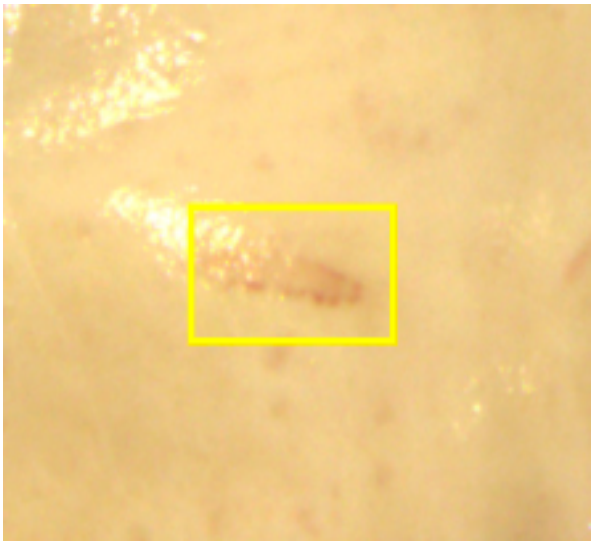


Abbildung 7: Skin Uncertain

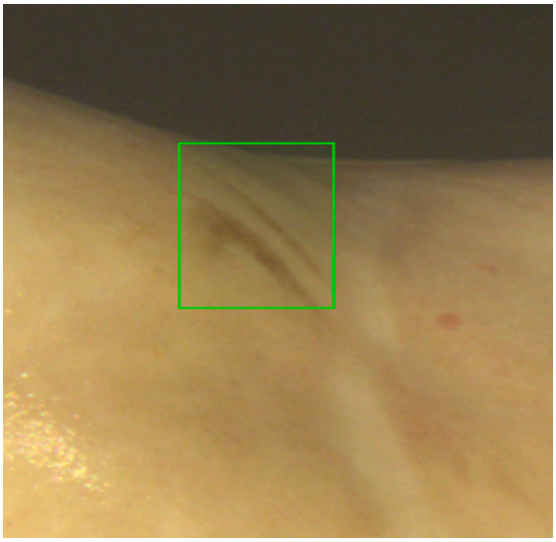


Abbildung 8: Skin Old Scratch

Tabelle 3: Labels und Definition der Schwanzläsionen zur Bildannotierung in CVAT

Bezeichnung des Labels	**Definition**
TailBiteMark	Einzelne oder gruppierte (zahnreihenförmige) Stichwunde(n) mit Anzeichen von Entzündung(en).
TailSuperficialWound	Verletzung der Haut, die keine tieferen Gewebsstrukturen wie Muskeln, Sehnen oder Knochen betrifft.
TailDeepWound	Verletzung mit Beteiligung von tieferen Gewebsstrukturen wie Muskeln, Sehnen oder Knochen.
TailBruise	Purpurrote/violette Färbung der Haut (verursacht durch Kapillarblutung)
TailNecrosis	Großflächiger Zelltod mit Entzündungsreaktion, welche sich durch dunkelbraune oder schwarze Hautverfärbungen zeigt.
TailProcessCaused	Veränderungen, die während des Prozesses des Schlachtens bzw. des Brüh-/ und Entborstungsvorganges entstehen.
TailUncertain	Für den Gutachter nicht definierbare Veränderung.
TailSwelling	Zunahme des Volumens eines Gewebes.
TailDisease	Durch einen Erkrankungsprozess verändertes Gewebe.

TailColourChangeTipWithTissueLoss	Rotverfärbung der Schwanzspitze mit Zusammenhangstrennung des Gewebes.
TailColourChangeTipWithoutTissueLoss	Rotverfärbung der Schwanzspitze ohne Zusammenhangstrennung des Gewebes.



Abbildung 9: Tail Color Change Tip Without Tissue Loss

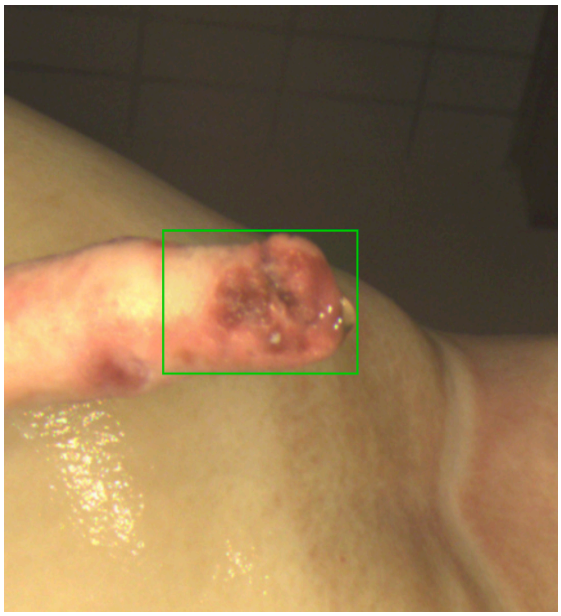


Abbildung 10: Tail Deep Wound

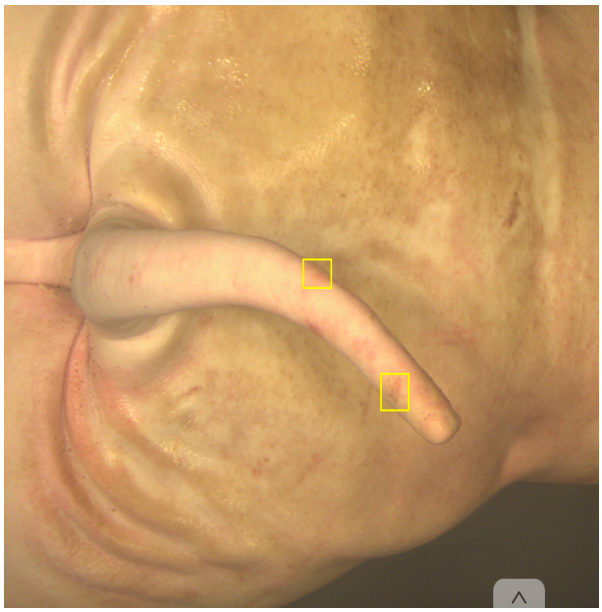


Abbildung 11: Tail Bite Mark

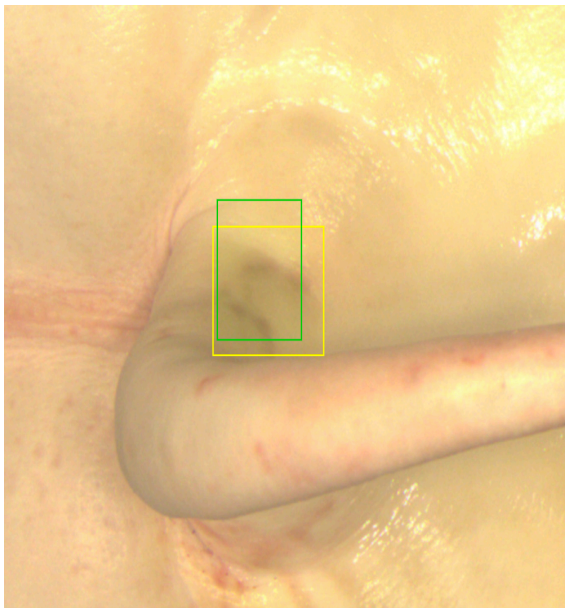


Abbildung 12: Tail Bruise

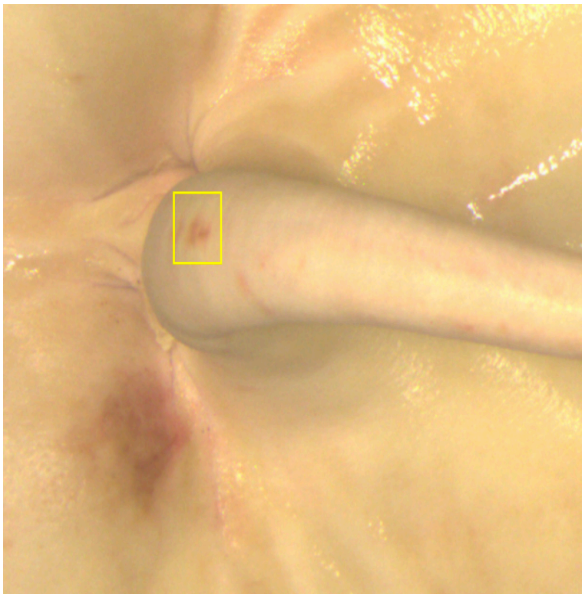


Abbildung 13: Tail Superficial Wound

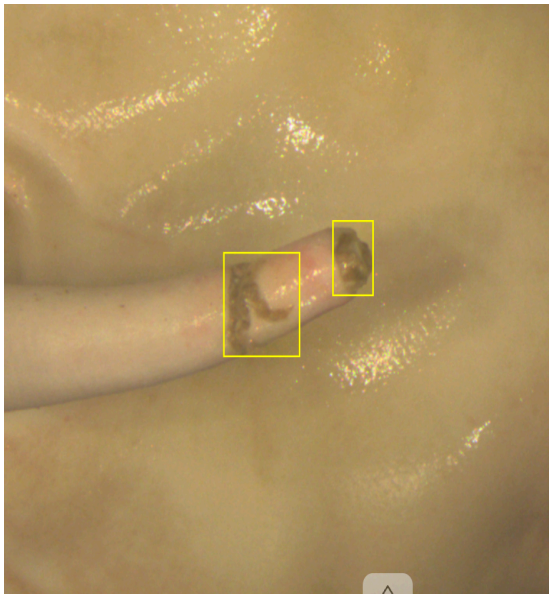


Abbildung 14: Tail Process Caused

Im späteren Verlauf der Annotierungen werden außerdem zusätzlich die Tätowierungen der Tiere mittels einer eigens Kategorie markiert. Zur eindeutigen Identifikation der Herkunftsbetriebe schreibt die Tierkennzeichnungs- und Registrierungsverordnung (TKZVO 2009) eine dauerhafte und gut lesbare Kennzeichnung von Schweinen vor. Die Tätowierung ist dabei so anzubringen, dass sich maximal zwei pro Schlachttier ergeben, entweder beidseitig im Schulterbereich oder beidseitig außen im Flankenbereich [6]. Mehr als zwei Markierungen pro Tier stellen - bei häufigem Vorkommen - einen tierschutzrelevanten Verstoß dar.

Um ein möglichst objektive und standardisierte Annotation zu gewährleisten wird außerdem ein bestimmter Ablauf für die Bildbeurteilung festgelegt:

- Zwei Beobachter benutzten zwei Geräten desselben Modells (Apple MacBook mit 13,3-Zoll-Display) um die Bilder unter identischen technischen Bedingungen zu beurteilen
- Das Heranzoomen der Bilder ist nicht gestattet
- Beide Laptops verfügen über identische Bildschirmdimensionen
- Während der Annotation wird auf konstante Lichtverhältnisse geachtet, um visuelle Abweichungen zu vermeiden

2.4 Durchführung der Beobachterabgleiche durch zwei beteiligte Beobachterinnen

Grundsätzlich dient ein Beobachterabgleich im Zuge der Algorithmenentwicklung dazu, die Objektivität der Annotationen durch verschiedene Personen zu überprüfen. Zudem ermöglicht ein Beobachterabgleich die Erstellung einer validen *Ground Truth,* welche wiederum als objektive Referenz für den Algorithmus dient.

Die *Ground Truth,* auch Goldstandard genannt, bedeutet in der Forschung die korrekten oder tatsächlichen Daten, mit welchen ein bestimmtes Modell oder Algorithmus verglichen werden soll. Sie dient als wichtige Referenz zur Beurteilung der Genauigkeit maschinellen Lernens [7].

Das Ziel der Beobachterabgleiche ist es, herauszufinden, inwieweit zwei Beurteilerinnen in der Vergabe der unterschiedlichen Labels für identische Bildsets übereinstimmen. Dafür werden die Datensätze vom 23. und 30. Juli 2024 herangezogen und in insgesamt drei Durchgängen beurteilt.

Die Bilder werden dabei zunächst unabhängig vom jeweils anderen Beobachter bewertet. Das bedeutet, dass es bei dem ersten Beobachterabgleich vorab zu keinem internen Austausch zwischen beiden zu beurteilenden Personen gekommen ist.

Im Zuge des zweiten Annotierungsdurchganges findet zuvor ein fachlicher Austausch zwischen beiden beteiligten Beobachterinnen mit dem Ziel, die Übereinstimmung zu erhöhen, statt. Es werden zunächst die verschiedenen Arten der Hautläsionen detailliert besprochen und deren Erscheinungsbild definiert. Außerdem wird auch diskutiert, welche Läsionen als relevant für die Annotierungen gelten und welche Veränderungen nicht weiter annotiert werden sollen, da sie vermutlich mit keiner relevanten Beeinträchtigung des Tierwohls verbunden sind.

Auch vor dem dritten und letzten Beobachterabgleich findet eine Diskussion statt, um die Positionsgenauigkeit der Rechtecke zu erhöhen, mit welchen die Läsionen an den Tierkörpern markiert werden.

In den beiden Besprechungen zwischen den Beobachterinnen liegt der Schwerpunkt hauptsächlich auf Anpassung eines Klassifikationssystems für die unterschiedlichen Läsionskategorien. Es werden auch die genauen Kriterien, welche eine Läsion aufweisen muss, um in die jeweilige Kategorie geordnet zu werden, nochmals besprochen.

Außerdem wird auch die Anzahl der zu vergebenen Rechtecke pro Bild diskutiert.

2.5 Statistische Auswertung der Daten

Für die Berechnung der Übereinstimmung bei den Beobachterabgleichen werden gewichtete und ungewichtete Kappa-Koeffizienten und ihre 95%-Konfidenzintervalle berechnet. Auch für den Vergleich der manuellen Erhebung von Hautläsionen direkt am Schlachthof mit der bildbasierten Annotierung mittels CVAT wird der Kappa-Koeffizient berechnet.
Dieses statistische Maß lässt sich mit folgender Formel berechnen:

$$\kappa = \frac{Po - Pe}{1 - Pe}$$

κ = Cohens Kappa
Po= beobachtete Übereinstimmung
Pe= erwartete Übereinstimmung

Cohens Kappa ist ein Maß zur Bewertung der Interrater- oder Intrarater Reliability. Interrater Reliability beschreibt, wie gut mehrere Personen (Rater) bei der Bewertung übereinstimmen. Intrarater Reliability bezieht sich hingegen darauf, wie zuverlässig eine einzelne Person bei wiederholter Bewertung derselben Situation ist [8].
Als Korrelationskoeffizient liegt der Wert von Kappa zwischen -1 und +1, wobei 0 eine Übereinstimmung darstellt, die als zufällig gilt und 1 eine perfekte Übereinstimmung signalisiert [8].
Die Kappa-Werte lassen sich wie folgt interpretieren [9]:

- ≤ 0: keine Übereinstimmung
- 0,01–0,20: geringe Übereinstimmung
- 0,21–0,40: mäßige Übereinstimmung
- 0,41–0,60: mittlere Übereinstimmung
- 0,61–0,80: substanzielle Übereinstimmung
- 0,81–1,00: nahezu perfekte Übereinstimmung

In der Fachliteratur wird dabei ein Mindestwert von 80 % Übereinstimmung als akzeptable Interrater Reliablity angenommen [8].
Kappa Werte unter 0,60 weisen auf unzureichende Übereinstimmung hin und sollten somit als problematisch beurteilt werde. Hier ist die Aussagekraft der Ergebnisse in Frage zu stellen.
Zur Bewertung der Übereinstimmung der Beobachterabgleiche vom 23. und 30. Juli 2024 werden durch CLK mehrere Klassifikationsmetriken berechnet und grafisch dargestellt. Pro Termin werden je drei Iterationen durch Beobachterin A und Beobachterin B durchgeführt.

In der Auswertung werden verschiedene Kennzahlen für jede Iteration herangezogen, um die Qualität der Übereinstimmung zu beurteilen und zusätzlich sogenannte Konfusionsmatrizen erstellt.

Als Grundlage für die Bewertung wird dabei eine Beobachterin (Beobachterin A) als *Ground Truth* definiert. Die Konfusionsmatrizen werden durch CLK auf Basis aller annotierten Bilder der beiden Beobachterabgleiche generiert, wobei doppelt vorhandene Aufnahmen desselben Schweines aus der Auswertung entfernt werden. Die Matrizen bestehen aus den folgenden vier Kategorien:

- True Positives (TP): beide Beobachterinnen haben dieselbe Läsionskategorie markiert
- False Positives (FP): nur Beobachterin B hat eine bestimmte Kategorie annotiert
- False Negatives (FN): Nur Beobachterin A, in diesem Fall die *Ground Truth*, hat diese Kategorie verwendet
- True Negatives (TN): keine der beiden Beobachterinnen hat diese Kategorie verwendet

Nur wenn beide Beobachterinnen auf dem gleichen Bild die gleiche Läsionskategorie z.B. „SkinBruise“ annotieren, unabhängig von der Lokalisation am Bild, wird dies als True Positive (TP) angezeigt. Wichtig zu erwähnen ist außerdem, dass die Matrizen nicht die Leistung eines einzelnen Klassifikators darstellen, sondern viel mehr für den Vergleich von zwei BeobachterInnen genutzt werden (Alexander Voß, persönliche Mitteilung, 17.06.2025).

Die statistischen Kennzahlen, welche zur Berechnung der Übereinstimmung herangezogen wird, basieren auf den erstellten Konfusionsmatrizen und beinhalten neben dem bereits erwähnten Cohens Kappa auch andere Kennzahlen, welche in Tabelle 4 dargestellt werden.

Tabelle 4: Formeln und Definitionen der Kennzahlen zur Erstellung der Konfusionsmatrix [10]

Kennzahl	**Formel**	**Definition**
Genauigkeit	(TP+TN)/(TP+FP+TN+FN)	Anteil der korrekt klassifizierten Fälle
Präzision	TP/(TP+FP)	Anteil der korrekt als positiv klassifizierten Fälle

Sensitivität	TP/(TP+FN)	Anteil der tatsächlich positiven Fälle, die korrekt erkannt wurden
Spezifität	TN/(TN+FP)	Anteil der korrekt als negativ klassifizierten Fälle

TP.....True Positives

TN.....True Negatives

FP.....False Positives

FN.....False Negatives

Diese vier Metriken werden berechnet, um die Annotierungen beider Beobachterinnen zu vergleichen. Außerdem ist es dadurch möglich, zu zeigen, ob eine bestimmte Beobachterin eher zu viel oder zu wenig markiert und ob sich die Übereinstimmung im Laufe der Durchgänge verbessert. Sie sind demnach unerlässlich um die Zuverlässigkeit der annotierten Daten zu bewerten und zu zeigen.

Die berechneten Metriken werden anschließend durch CLK für jede der insgesamt sechs Iterationen (drei pro Termin) grafisch in Form von Diagrammen aufbereitet. Auf diese Weise kann die Entwicklung der Übereinstimmung zwischen beiden Beobachterinnen im Verlauf der Durchgänge nachvollzogen werden.

Für den methodischen Vergleich zwischen der manuellen Befundung am Schlachthof und der bildbasierten Beurteilung mittels CVAT werden die erhobenen Daten mit Microsoft Excel aufbereitet. Im Rahmen der deskriptiven Statistik erfolgt die Berechnung absoluter Häufigkeiten der verschiedenen Läsionskategorien und aufgezeichneten Lokalisationen, um Unterschiede zwischen beiden Erhebungsmethoden darstellen zu können.

2.6 Optimierung der Ground Truth im Expertenteam

Im Rahmen dieser Arbeit ist es im Verlauf der Algorithmenentwicklung vorgesehen, innerhalb einer Expertengruppe eine *Ground Truth* für die Annotierung von Bildmaterial zu etablieren. Die Definition der *Ground Truth* ist ein essentieller Bestandteil bei der Erstellung eines Algorithmus und zu ihrer Entwicklung waren mehrere Schritte notwendig.

Das Expertenteam besteht aus fünf Personen mit veterinärmedizinischem Hintergrund und Erfahrung bei der Befundung von Schlachtkörpern.

Zunächst muss festgelegt werden, was überhaupt beurteilt oder annotiert werden soll. Außerdem muss auch definiert werden, welche morphologischen Merkmale eine Läsion aufzuweisen hat, um in eine bestimmte Kategorie geordnet zu werden und ab wann sie überhaupt als relevant gilt.

Das Ziel soll es sein, nachvollziehbare und standardisierte Kriterien zu bestimmen um eine einheitliche Arbeitsweise zu gewährleisten.

Zur Erstellung der *Ground Truth* wird im zweiten Schritt Bildaufnahmen aus verschiedenen Erhebungszeitpunkten gemeinschaftlich im Team analysiert. Die Annotationen werden wieder direkt mit Hilfe von CVAT durchgeführt. Die ExpertInnen führen die Bewertung gemeinsam vor einem Computerbildschirm durch, wobei zunächst jeweils ein Experte beschreibt, welche Läsionen wahrgenommen wird. Anschließend bestätigen oder widersprechen die übrigen Experten jeweils diese Einschätzung.

Danach wird die identifizierten Läsionen ebenfalls wieder mit einem Rechteck annotiert und den jeweiligen Kategorien zugeordnet.

Bei der ersten Sitzung werden dabei sämtliche sichtbare Läsionen auf den Schlachtkörpern markiert, unabhängig davon ob diese prozessbedingt postmortal oder intra vitam entstanden sind.

Prozessbedingte Läsionen sind jene, welche durch den Schlachtprozess selbst entstehen, beispielsweise durch mechanische Einwirkungen während verschiedener Prozesse.

Gemäß österreichischem Tierschutzgesetz §32 Abs. 1 gilt eine intra vitam entstandene Verletzung dann als tierschutzrelevant, wenn sie auf vermeidbare Schmerzen oder Leiden hindeutet und dem Tier vor oder während der Schlachtung zugefügt wurde [11].

Tierschutzrelevante Läsionen sind demnach Verletzungen, welche – wenn sie erheblich und langanhaltend sind - auf Verstöße gegen das Tierschutzgesetz hinweisen. Diese entstehen meist durch falschen Umgang oder mangelhafte Haltung im Herkunftsbetrieb, während des Transports und oder im Wartebereich und beim Zutrieb zur Betäubung im Schlachtbetrieb.

In der Praxis fällt der Expertengruppe die Differenzierung zwischen vermeintlich tierschutzrelevanten und prozessbedingten Läsionen nach den Konsultationen zunehmend schwerer und es folgt eine schrittweise Reduktion auf die tatsächlich tierschutzrelevanten Läsionen. Dieser Entschluss führt in den darauffolgenden Sitzungen zu einer deutlichen Reduktion der Markierungen pro Bild.

3. Ergebnisse

Der Ergebnisteil der Arbeit umfasst insgesamt drei Teile. Der erste Teil beschäftigt sich mit den Darstellungen der Prävalenzen der erhobenen Hautläsionen direkt am Schlachthof. Der zweite Teil beschreibt den Vergleich der erhobenen Hautläsionen am Schlachthof im Vergleich mit der bildbasierten Bewertung mittels CVAT. Im dritten Teil werden abschließend die Ergebnisse der Beobachterabgleiche mithilfe statistischer Kennzahlen dargestellt.

3.1 Prävalenzen von Hautläsionen erhoben am Schlachthof Higelsberger

Im Rahmen dieser Arbeit werden insgesamt 763 Mastschweine hinsichtlich des Auftretens von Hautläsionen direkt am Schlachthof Higelsberger in Schwertberg beurteilt. Die Erhebungen werden mittels dem Protokoll zur Datenerhebung am Schlachthof aufgenommen und anschließend in Excel aufbereitet.
In Tabelle 5 sind die Anzahl sowie die berechneten Prävalenzen in % dargestellt.

Tabelle 5: Absolute Häufigkeiten und Prävalenzen der Hautläsionen der untersuchten Schlachtschweine (n=763)

Läsionskategorie	Anzahl	Prävalenz (%)
SkinBruise	278	36,39
SkinFreshScratch	206	26,96
SkinOldScratch	120	15,71
SkinUncertain	53	6,94
SkinDisease	7	0,92
SkinOldDeepwound	2	0,26
SkinFreshDeepwound	1	0,13
NA	96	12,57

Obwohl das Erhebungsprotokoll eine Vielzahl verschiedenster Läsionskategorien vorsieht, bringt die tatsächliche Erhebung am Schlachthof deutlich weniger Abweichungen zu Tage.
Bisswunden werden beispielsweise nie dokumentiert und auch Kategorien wie „human handling“ werden kaum verwendet. Auch schwere Verletzungen wie tiefe Wunden treten mit einer einzigen Dokumentation nur selten auf.
Die am häufigsten festgestellten Läsionen sind Prellungen (SkinBruise) (36,39 %) und frische Kratzer (SkinFreshScratch) (26,96 %).

3.2 Prävalenzen von manuell am Schlachthof erhobenen Hautläsionen im Vergleich mit bildbasierter Bewertung

Zum Vergleich der Annotierungen, welche einerseits bildbasiert durch CVAT, andererseits manuell am Schlachthof erhoben werden, werden drei Erhebungstermine (06.08, 13.08. und 27.08.2024) herangezogen. Insgesamt werden 250 Tiere beurteilt.

Ziel ist es, die Übereinstimmung beider Methoden im Hinblick auf Lokalisation der Läsionen sowie deren Kategorisierung als tierschutzrelevant oder prozessbedingt zu analysieren.

Zur Auswertung wird aus den erhobenen Datensätzen eine Excel-Tabelle erstellt. In dieser werden die Annotierungen beider Verfahren für jedes Schwein gegenübergestellt. Dabei wird jeweils zuerst die Beobachtung am Schlachtband dokumentiert (Lokalisation und Läsionstyp) und anschließend überprüft, ob dieselbe Veränderung auch bildbasiert sichtbar und identisch auftritt. Bei Abweichungen wird festgehalten, ob die Läsion anders kategorisiert ist, oder eventuell gar nicht sichtbar ist. Diese Vorgehensweise ermöglicht eine direkte Vergleichbarkeit zwischen beiden Methoden auf Einzeltier-Ebene.

In Tabelle 6 werden die Nennungen beziehungsweise die relativen Häufigkeiten der Lokalisationen der Läsionen dargestellt, wobei NA bedeutet, dass keine Läsion dokumentiert wird. Zur Bestimmung der jeweiligen Lokalisationen einer Läsion wird das in Abbildung 15 dargestellte Schema angewendet.

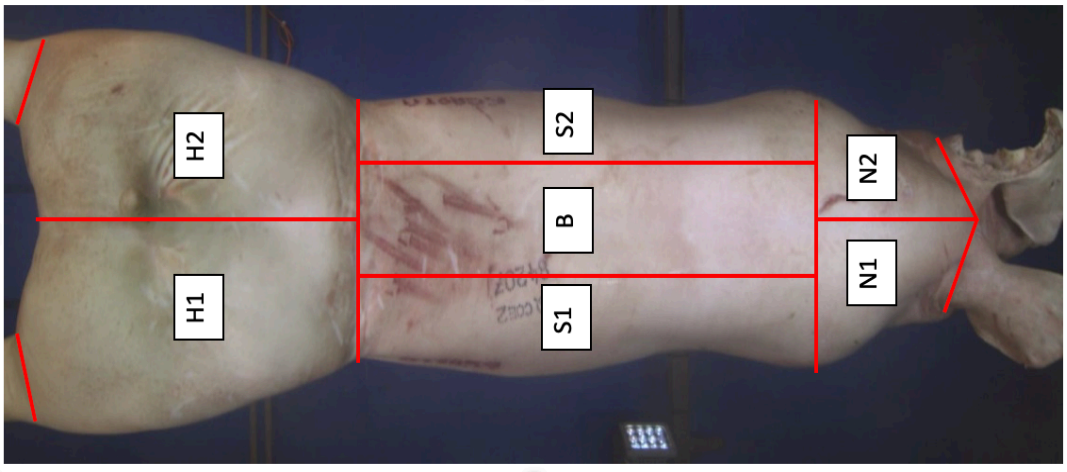


Abbildung 15: Schematische Darstellung der Lokalisationen am Schlachtkörper

Tabelle 6: Relative Häufigkeiten [%] der Lokalisation von Läsionen im Vergleich von Bildbewertung (CIVAT) und direkter Bewertung am Schlachtband

	B	H1	H2	N1	N2	S1	S2	NA
CVAT	50	18	23	4	2	13	9	131
Schlachtband	73	36	32	25	17	19	16	32

Diese Ergebnisse zeigen, dass die direkte Erhebung am Schlachthof tendenziell mehr Lokalisationen erfasst, insbesondere im Bereich des Nackens (N1 und N2).
Auch bei den Läsionstypen gibt es Unterschiede zwischen beiden Methoden. Diese werden in Tabelle 7 dargestellt. Die Definition der Läsionen im Rahmen dieser Analyse orientiert sich an jener zur Erfassung von Hautläsionen (Tabelle 1).

Tabelle 7: Absolute Häufigkeiten der Läsionstypen je Erfassungsmethode

	Bruise	Disease	Fresh Scratch	Old Scratch	Uncertain	Not Available
CVAT	52	-	18	20	22	131
Schlachtband	91	5	63	30	29	32

Außerdem werden die Läsionen auch hinsichtlich der Tierschutzrelevanz untersucht. Im Gegensatz dazu stehen Läsionen, welche durch prozessbedingte Vorgänge entstehen. Die Ergebnisse dieser Untersuchung sind in Tabelle 8 dargestellt.

Tabelle 8: Läsionen annotiert als tierschutzrelevant oder prozessbedingt

	nein	ja	NA
tierschutzrelevant	87	32	131
prozessbedingt	32	88	130

Bei der Bildannotation mittels CVAT werden demnach die Läsionen eher als prozessbedingt einzustufen, während die direkte Beurteilung am Schlachtband häufiger tierschutzrelevante Befunde erkannte.
Bei den Kappa Werten der Lokalisationen ergibt sich ein ungewichteter Kappa Wert von κ= 0,99 (Konfidenzintervall 0,96-1,00) und ein gewichteter Kappa-Wert von κ= 1,00 (Konfidenzintervall: 0,99-1,00).

Dies weist auf eine nahezu perfekte Übereinstimmung hin, was daran liegt, dass nur jeweils dieselbe Lokalisation beurteilt wurde und somit keine Differenzen zu erwarten sind. Das Konfidenzintervall liegt bei beiden Werten zudem weit über 0 und zeigt somit, dass die Übereinstimmung signifikant und kein statistischer Zufall ist.
Bei den Läsionstypen hingegen zeigt sich ein ungewichteter Kappa-Koeffizient von κ= 0,5 (Konfidenzintervall: 0,38-0,62), was gemäß der Klassifikation auf eine moderate Übereinstimmung hinweist [9].
Der gewichtete Kappa Koeffizient beträgt κ=0,55 (Konfidenzintervall 0,39-0,70) und liegt damit ebenfalls im Bereich der moderaten Übereinstimmung [9]. Hier ist das Konfidenzintervall jedoch breiter, was auf eine größere Unsicherheit bei der Übereinstimmung hinweist. Allerdings liegt auch hier das Intervall vollständig über 0 und somit ist auch diese Übereinstimmung statistisch signifikant, wenn auch in geringerem Ausmaß [8].

3.3 Beobachterabgleich zwischen zwei beteiligten Beobachterinnen

Um die Fehlstellen (Läsionen) an den Schlachtkörpern korrekt zu identifizieren, werden vorab Labels definiert, mit denen die Fehlstellen im Bild als Rechteck annotiert werden (Alexander Voß, persönliche Mitteilung, 08.04.2025). Das Ziel ist es, herauszufinden, inwieweit zwei Beobachterinnen in der Beurteilung der unterschiedlichen Labels übereinstimmen.
Die Bilder werden zunächst unabhängig zu der jeweils anderen Beobachterin bewertet. Da die Ergebnisse des ersten Durchganges jedoch nicht zufriedenstellend waren, findet im Zuge der zweiten Annotierungsrunde ein Gespräch zwischen beiden beteiligten Beobachterinnen statt. Das Ziel ist es, die Übereinstimmung zu erhöhen.
Auch bei der dritten und letzten Annotierung findet ebenfalls vorab eine Interaktion statt, um die Positionsgenauigkeit der Rechtecke zu erhöhen, mit welchen die Fehlerstellen markiert werden. Zur Beurteilung der Übereinstimmung werden die Datensätze vom 23. und 30. Juli 2024 herangezogen. Dabei werden pro Termin jeweils drei Iterationen durchgeführt und die Ergebnisse anschließend durch die CLK GmbH grafisch dargestellt.

3.3.1 Ergebnisse der Beobachterabgleiche vom 23.07.2024 und 30.07.2024

Im ersten Beobachterabgleich werden drei aufeinanderfolgende Beobachtungsdurchgänge (Iterationen 1, 2 und 3) durchgeführt. Die Bewertung erfolgt auf Basis der Bilddateien im CVAT System. Folgende Tabellen 9 und 10 zeigen jeweils die berechneten Metriken für jede verwendete Läsionskategorie pro Versuch.

Tabelle 9: Ergebnisse des Beobachterabgleiches vom 23.07.2024

Läsionskategorie	Genauigkeit 1	Präzision 1	Sensitvity 1	Spezifität 1	Cohens Kappa 1	Genauigkeit 2	Präzsion 2	Sensitivität 2	Spezifität 2	Cohens Kappa 2	Genauigkeit 3	Präzision 3	Sensifivity 3	Spezifität 3	Cohens Kappa 3
SkinBruise	0,71	0,79	0,61	0,82	0,43	0,94	1,0	0,89	1,0	0,88	0,82	0,85	0,92	0,55	0,5
SkinFreshScratch	0,94	0,29	1,0	0,94	0,42	0,93	0,69	0,82	0,95	0,71	0,82	0,5	0,6	0,87	0,44
SkinOldScratch	0,75	1,0	0,05	1,0	0,07	0,92	0,88	0,85	0,95	0,8	0,74	0,68	0,85	0,64	0,48
SkinUncertain	0,9	0,0	0,0	0,97	-0,04	1,0	1,0	1,0	1,0	1,0	0,79	0,62	0,89	0,73	0,56
TailBiteMark	-	-	-	-	-	1,0	1,0	1,0	1,0	1,0	0,82	0,62	0,29	0,96	0,31
TailBruise	-	-	-	-	-	0,99	1,0	0,83	1,0	0,9	0,87	0,8	0,6	0,95	0,61
TailColourChangeTipWithoutTissueLoss	0,95	0,2	0,1	0,95	0,32	1,0	1,0	1,0	1,0	1,0	-	-	-	-	-
TailColourChangeWithTissueLoss	-	-	-	-	-	-	-	-	-	-	0,93	1,0	1,0	0,14	0,23
TailColourChangeWithoutTissueLoss	-	-	-	-	-	-	-	-	-	-	0,74	0,25	0,88	0,72	0,28
TailDeepWound	0,99	0,5	1,0	0,99	0,66	-	-	-	-	-	-	-	-	-	-
TailProcessCaused	-	-	-	-	-	0,99	1,0	0,75	1,0	0,85	0,98	1,0	0,5	1,0	0,66
TailSuperficialWound	-	-	-	-	-	0,96	0,6	0,75	0,98	0,65	0,88	0,36	0,57	0,91	0,38
TailUncertain	-	-	-	-	-	-	-	-	-	-	0,9	0,14	0,33	0,93	0,16

Tabelle 10: Ergebnisse des Beobachterabgleichs vom 30.07.2024

Läsionskategorie	Genauigkeit 1	Präzsion 1	Sensitivität 1	Spezifität 1	Cohen's Kappa1	Genauigkeit 2	Genauigkeit 2	Sensitivität 2	Spezifität 2	Cohen's Kappa2	Genauigkeit 3	Präzision 3	Sensitivität 3	Spezifität 3	Cohen's Kappa 3
SkinBruise	0,69	0,96	0,53	0,97	0,43	0,85	0,95	0,83	0,9	0,67	0,89	0,98	0,89	0,93	0,69
SkinDisease	0,99	0,0	0,0	0,99	-0,01	-	-	-	-	-	-	-	-	-	-
SkinFreshScratch	0,95	0,36	0,4	0,97	0,35	0,79	0,13	0,7	0,8	0,16	0,8	0,31	0,65	0,82	0,32
SkinOldScratch	0,73	0,94	0,21	0,99	0,25	0,72	0,51	0,72	0,72	0,39	0,8	0,76	0,88	0,72	0,59
SkinUncertain	0,96	0,67	0,5	0,99	0,55	0,97	0,58	0,88	0,98	0,69	0,88	0,62	0,75	0,91	0,61
TailBiteMark	-	-	-	-	-	0,95	0,75	0,5	0,99	0,57	0,92	0,7	0,82	0,94	0,71
TailBruise	-	-	-	-	-	0,92	0,53	0,47	0,96	0,46	0,87	0,69	0,6	0,94	0,57
TailColourChangeTipWithTissueLoss	-	-	-	-	-	1,0	1,0	1,0	1,0	1,0	0,99	0,33	1,0	0,99	0,5
TailColourChangeTipWithoutTissueLoss	0,97	0,36	0,8	0,97	0,49	1,0	1,0	0,92	1,0	0,96	0,99	0,33	1,0	0,99	0,5
TailDeepWound	1,0	0,5	1,0	1,0	0,66	1,0	1,0	1,0	1,0	1,0	1,0	1,0	0,5	1,0	0,66
TailProcessCaused	-	-	-	-	-	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0
TailSuperficialWound	0,99	0,5	0,5	1,0	0,5	0,92	0,27	0,92	0,88	0,38	0,88	0,24	0,8	0,89	0,32
TailUncertain	-	-	-	-	-	-	-	-	-	-	0,98	0,83	0,56	1,0	0,6

Die Ergebnisse der ersten Iteration zeigen eine geringe Übereinstimmung zwischen den Beobachterinnen hinsichtlich der gewählten Läsionskategorien. Die Genauigkeitswerte liegen durchweg hoch (zwischen 0,71 und 0,99), was jedoch kein aussagekräftiges Maß für Übereinstimmung darstellt [10].
Besonders auffällig ist die niedrige Sensitivität bei „SkinOldScratch“ und niedrige Genauigkeit bei „SkinFreshScratch“ und „TailColourChangeTipWithTissueLoss“, was auf eine uneinheitliche Anwendung dieser Kategorien hinweist. Beobachterin B hat demnach diese Art der Läsionen häufig verwendet, obwohl Beobachterin A sie nicht gesehen hat.
Die Übereinstimmung ist insgesamt in diesem Durchgang gering bis moderat und besonders die Kategorien wie „SkinUncertain“ und „SkinOldScratch“ weisen große Differenzen auf.
Vor dem zweiten Versuch findet ein persönlicher Austausch zwischen beiden Beobachterinnen, mit dem Ziel die Übereinstimmung zu erhöhen, statt. Cohen’s Kappa liegt hier fast durchgehend über 0,80, was auf eine gute Übereinstimmung hinweist [9]. Die Genauigkeits-Werte liegen für alle Kategorien zwischen 0,92 und 1,0 was auf einen hohen Anteil an korrekt klassifizierten Fällen hinweist. Im zweiten Versuch zeigt sich also eine deutliche Verbesserung in allen Metriken und fast allen Kategorien.
Die hohe Spezifität (≥ 0,95 in fast allen Kategorien) weist darauf hin, dass gesunde Hautbereiche zuverlässig nicht als Läsion markiert wurden. Dies ist wichtig, um Fehlklassifikationen zu vermeiden [10]. Die Schwächen bei Kategorien wie „SkinOldScratch“ und „SkinUncertain“ aus der ersten Runde werden deutliche reduziert.
Die Ergebnisse des dritten Versuches zeigen, dass die Übereinstimmung zwischen den Beobachterinnen im Vergleich zu zweiten Iteration schwächer ausfällt. Die Genauigkeits-Werte liegen nun zwischen 0,74 und 0,98 und auch deutlich niedrigere Kappa Werte (0,16-0,66) wird in dieser Runde berechnet. Dies deutet auf Uneinigkeit zwischen den Beobachterinnen bei der Interpretation einzelner Kategorien hin. Besonders Labels wie „SkinUncertain“ oder „TailBiteMark“ verzeichnen deutlich geringe Übereinstimmung.

Insgesamt lässt sich festhalten, dass zwischen der ersten und dritten Wiederholung starke Verbesserungen in der Beobachterinnen-Übereinstimmung erreicht werden. Besonders deutlich ist der Anstieg bei der Genauigkeit. Diese steigt etwa bei „SkinBruise“ von 0,71 auf 0,82.
Die Genauigkeit, also der Anteil an tatsächlich korrekter positiver Annotationen durch Beobachterin B, ist in der ersten Wiederholung noch uneinheitlich. Beispielsweise liegt sie bei

„SkinFreshScratch“ bei 0,29. In der dritten Wiederholung jedoch bei 0,50, was auf eine deutliche Verbesserung hinweist.
Auch Cohens Kappa verbessert sich signifikant. Beispielhaft kann hier ein Anstieg bei Labels wie „SkinOldScratch“ von 0,07 auf 0,48 und für „SkinUncertain“ von -0,04 auf 0,56 verzeichnet werden.
Im zweiten Beobachterabgleich vom 30.07.2024 werden wiederum drei Durchgänge durchgeführt.
Auch in der ersten Annotierungsrunde dieses Beobachterabgleiches erfolgt zuvor keine Absprache zwischen den Beobachterinnen. Dies spiegelt sich in den berechneten Kappa Werten wider, welche insgesamt als niedrig bis moderat zu beurteilen sind. Besonders problematisch sind Kategorien wie „SkinDisease“ und „SkinFreshScratch“ mit wenig Übereinstimmung. Die Genauigkeits-Werte sind durchwegs hoch, allerdings kann dies irreführend sein und es ist daher immer die Kombination mit Genauigkeit, Sensitivität und Kappa essenziell [10].
In diesem Durchgang (oder Wiederholung) zeigt sich ein deutlicher Fortschritt gegenüber der ersten Runde. Besonders bei Kategorien wie „SkinBruise“ oder „TailDeepWound“ wird eine gute Übereinstimmung erzielt. Die durchgehend hohen Genauigkeiten deuten darauf hin, dass sich die Beobachterinnen häufig über das Vorhandensein (True Positive) oder nicht vorhanden sein (True Negative) einer Läsion einig sind. Die hohen Präzisionen bei Labels wie beispielsweise „SkinBruise“ zeigen, dass die Markierungen von Beobachterin B häufig korrekt waren. Niedrige Genauigkeits-Werte, wie sie hingegen bei beispielsweise „SkinFreshScratch“ berechnet wurden, bedeuten in diesem Fall, dass Beobachterin B diese Läsionen markiert hat, obwohl Beobachterin A diese nicht markiert hat. Insgesamt ist hier jedoch eine stark verbesserte Übereinstimmung im Vergleich zur ersten Wiederholung dieses Termines zu verzeichnen.
In der dritten Annotierungsrunde wird zusätzlich im Vorfeld über die exakte Platzierung der Rechtecke diskutiert, um die Konsistenz der Annotationen weiter zu steigern.
In dieser Wiederholung bleib die Gesamtqualität im Vergleich zur zweiten Wiederholung stabil hoch. Die durchwegs hohen Genauigkeiten bedeuten, dass beide Beobachterinnen in den meisten Fällen dieselbe Entscheidung getroffen haben. Hinsichtlich der Präzisionen zeigt sich ein uneinheitliches Bild. Bei manchen Labels wie „SkinBruise“ wurden hohe Werte berechnet. Bei Labels wie „TailSuperficialWound“ hingegen zeigt sich durch die niedrigere Präzision, dass diese Kategorie schwierig zu bewerten war.

Auch in diesem Beobachterabgleich lässt sich ein klarer Lernerfolg ableiten. Die Genauigkeits-Werte verbessern sich über die drei Iterationen hinweg systematisch. Beispielhaft steigt die Genauigkeit für „SkinBruise“ von 0,69 auf 0,89.
Besonders hervorzuheben ist der Zugewinn bei der Sensitivität. So steigt diese bei „SkinBruise“ von 0,53 auf 0,89 oder bei „SkinUncertain“ von 0,05 auf 0,75. Diese Entwicklung zeigt, dass Beobachterin B über die drei Wiederholungen hinweg mehr tatsächliche Läsionen korrekt erkennt und dabei auch seltener Läsionen übersieht. Auch bei Cohens Kappa kommt es zu einer Verbesserung, beispielsweise bei „SkinOldScratch“ von 0,25 auf 0,59. Diese Verbesserung zeigt, dass sich im Verlauf des Annotierungsprozesses deutlich mehr Übereinstimmung als zu Beginn erzielt werden können.

4. Diskussion

Die Diskussion dieser Arbeit gliedert sich in drei Teile. Zunächst werden die Herausforderungen bei der Klassifikation von Hautläsionen beschrieben und welche Läsionstypen sich als besonders schwierig zu identifizieren erwiesen und warum. Danach werden die Optimierungsschritte des Beobachterabgleiches dargestellt und diskutiert, wie die Übereinstimmungsrate innerhalb des Abgleiches noch weiter gesteigert werden kann.
Zuletzt werden noch die Schwierigkeiten beim Etablieren der *Ground Truth* im Expertenteam beschrieben und aufgezeigt, welche Schritte bei zukünftigen Projekten bessere Ergebnisse versprechen könnten.

4.1 Herausforderungen bei der Klassifikation von Hautläsionen als prozessbedingt oder tierschutzrelevant

Im Laufe des Annotierungsprozesses stellt sich heraus, dass es bei einem Großteil der Läsionen nicht möglich ist, eindeutig zu identifizieren, ob diese nun tatsächlich tierschutzrelevant sind oder nicht.
Zur Veranschaulichung, welche Läsionen jedoch als prozessbedingt oder tierschutzrelevant einzuordnen sind, dienen die folgenden Abbildungen.
Abbildungen 16 und 17 zeigen dabei als tierschutzrelevant zu bewertende Läsionen mit unterschiedlicher Herkunft. Abbildung 16 wurde mit einem Mobiltelefon direkt am Schlachtband aufgenommen. Die restlichen Abbildungen stammen hingegen direkt aus dem CVAT System.





Abbildung 16 (links): Tierschutzrelevante Läsionen (Kratzer) an der Flanke eines Schweineschlachtkörpers

Abbildung 16 (rechts): Tierschutzrelevante Läsion vermutlich durch unsachgemäße Anwendung des Elektroschockgerätes

Das linke Bild zeigt die Flanke eines Schweines mit zahlreichen parallel verlaufenden, teils überlappenden Kratzspuren. Die Lokalisation und das Muster der Läsionen sprechen stark für agonistisches Verhalten zwischen Artgenossen, welches insbesondere beim Mischen von verschiedenen Gruppen auftritt [12]. Abbildung 16 (rechts) hingegen zeigt mehrere kleine, rundliche Läsionen mit abgegrenzten dunklen Zentren. Die regelmäßige Anordnung könnte möglicherweise auf eine mechanische Ursache hindeuten, in diesem Fall vermutlich durch die unsachgemäße Anwendung einer Elektrotreibhilfe. Laut VO (EG) Nr. 1099/2009 über den Schutz von Tieren zum Zeitpunkt der Tötung ist die Verwendung von Elektroschockern nur in Ausnahmefällen erlaubt, nur an der Hinterhand und nicht länger als notwendig. Die sichtbaren Veränderungen lassen daher auf einen tierschutzrelevanten Verstoß schließen.

Abbildung 18 und 19 zeigen beide eine Prellung im Bereich des Rückens zweier Tiere.

Diese Läsionen können jedoch sowohl tierschutzrelevante (grober menschlicher Umgang, Verletzung beim Transport, ...), als auch prozessbedingte Ursachen haben.





Abbildung 17 (links): prozessbedingte Läsionen am Rücken eines Tieres

Abbildung 18 (rechts): prozessbedingte Läsionen am Rücken eines Tieres

Als mögliche Verletzungsquellen können hier mechanische Einflüsse während des Schlachtprozesses, aber auch Kontakt mit scharfen Kanten, harten Oberflächen und Prozesse wie das Abflämmen oder die Entborstung der Tiere genannt werden.

Es kann somit nicht immer mit eindeutiger Sicherheit klassifiziert werden, in welche Kategorie die betroffene Läsion schlussendlich einzuteilen ist.

Die Art und Lokalisation der Prellung wie in Abbildung 18 und 19 tritt allerdings häufig systematisch auf und deutet demnach tendenziell eher auf ein prozessbedingtes Geschehen, als auf tierschutzrelevante Verstöße hin. Die genaue Unterscheidung würde jedoch weitere Untersuchungen erforderlich machen.

Nur ein geringer Anteil an Läsionen ist im Rahmen dieser Arbeit eindeutig einer bestimmten Klassifikation zuzuordnen und viele Befunde bewegen sich in einem Graubereich. Während tierschutzrelevante Läsionen meist vermeidbares Leiden widerspiegeln, können prozessbedingte Läsionen durch unvermeidbare Abläufe während der Schlachtung entstehen und somit bleibt eine eindeutige Zuordnung oft schwierig [13].

Um tierschutzrelevante Läsionen in Zukunft zuverlässiger zu erkennen und zu vermeiden, wäre es denkbar, das sogenannte Animal Welfare Protokoll, insbesondere die Five Freedoms, hinzuzuziehen. Diese bieten eine praxisnahe Struktur zur direkten Ableitung von

Tierwohlindikatoren und es besteht die Möglichkeit, körperliche Befunde wie Wunden oder Verletzungen einzuordnen.
Das „Welfare Quality Assessment Protocol for Pigs“ formuliert die „Five Freedoms“ wie folgt [14]:
Freedom:

1. from hunger and thirst
2. from discomfort
3. from pain, injury and disease
4. to express normal behaviour
5. from fear and distress

Die Freiheit von Schmerz, Verletzungen und Krankheit (Freedom 3) steht dabei im direkten Zusammenhang mit dem Erkennen von tierschutzrelevanten Läsionen. Ebenso kann die Freiheit, normales Verhalten auszuleben (Freedom 4) herangezogen werden, um z.B. die Folgen agonistischen Verhaltens, wie beispielsweise Schwanzbeißen, zu interpretieren. Auch Furcht und Stress (Freedom 5) lassen sich indirekt über z.B. Verletzungen durch Fluchtverhalten nachvollziehen.
Langfristig gesehen könnten Sensor-basiert und KI-gestützte automatische Monitoring-Systeme am Schlachthof also dazu beitragen, solche Läsionen zu erkennen und zu klassifizieren.
Dass dieser Ansatz prinzipiell Potential hat, beweisen auch verschieden Projekte. So kann eine Studie beispielsweise zeigen, dass ein automatisiertes Kamerasystem zur Erkennung von Ohr und Schwanzverletzungen nach der Entwicklungsphase zum täglichen Einsatz am Schlachthof geeignet ist [1]. Die Autoren kommen darüber hinaus zu dem Schluss, dass dieses System auch zur Erkennung von Hautläsionen durch unsachgemäßen Umgang durch Menschen einsetzbar wäre [1].
Trotz dieser vielversprechenden Ergebnisse von automatisierten Systemen zur Erkennung von Läsionen zeigt sich, dass der Weg zu einer zuverlässigen und massentauglichen Anwendung noch lang sein wird. Herausforderungen bestehen vor allem in der Sammlung ausreichend großer Datensätze sowie der Integration in bereits bestehenden Datenbanken zur Kontrolle der Lebensmittelkette einschließlich des Tierschutzes, wie beispielsweise der Schlachttier- und Fleischuntersuchung, dem Fleischklassifizierungssystem der ÖFK oder dem Verbrauchergesundheitsinformationssystem VIS (https://vis.statistik.at/vis/schweine/haeufig-gestellte-fragen).

4.2 Optimierungsschritte des Beobachterabgleiches zwischen zwei Bewerterinnen und Auswirkungen auf die berechnete Übereinstimmung

Im Rahmen der Arbeit werden zwei Datensätze insgesamt dreimal unabhängig voneinander beurteilt, wobei sich die Ergebnisse im Laufe des Prozesses deutlich verbessert haben.
Grundsätzlich weist eine hohe Übereinstimmung der Bewertungen innerhalb des Beobachterabgleiches auf eine objektive Beurteilung der Fehlstellen an den Schlachtkörpern hin. Die Annotationen sind demnach nachvollzieh- und wiederholbar.
Eine geringere Übereinstimmung hingegen zeigt, dass Unsicherheiten bei der Lokalisierung der Läsionen sowie der korrekten Vergabe der verschiedenen Labels bestehen.
Die initial geringe Übereinstimmung ist unter anderem vermutlich auf unterschiedliche Interpretation der Labels sowie auf Missverständnisse darüber, wie viele Labels pro Tier verteilt werden sollen, zurückzuführen.
Nach dem ersten Durchgang findet ein gezielter Austausch statt und die beiden Beobachterinnen sprechen gemeinsam über die Bewertungskriterien anhand konkreter Beispiele.
Im zweiten Durchgang zeigt sich so bereits eine höhere Übereinstimmung, welche im dritten Durchgang nochmal erhöht werden kann.
Dies lässt sich deutlich an der Entwicklung der verschiedenen Metriken nachvollziehen. Im ersten Durchlauf des Beobachterabgleiches vom 23.07.2024 liegt der durchschnittliche Kappa Wert bei 0,29. Im dritten Durchlauf hingegen liegt der durchschnittliche Wert bei 0,47 mit hohen Werten in vielen Kategorien. Ebenso zeigt sich eine deutliche Steigerung bei der Sensitivität. Diese misst, wie viele der tatsächlich vorhandenen Läsionen korrekt erkannt werden und steigt von 0.05 auf 0,85. Das bedeutet, dass Beobachterin B in den späteren Iterationen deutlich mehr Läsionen erkennt, welche Beobachterin A ebenfalls annotiert [10].
Auch im ersten Durchgang des Beobachterabgleiches vom 30.07.2024 zeigt sich eine vergleichbare Entwicklung. Der durchschnittliche Kappa Wert liegt zunächst bei 0,40 mit besonders niedrigen Werten in bestimmten Kategorien. Im dritten Durchgang steigt der durchschnittliche Wert auf 0,59 mit teilweise moderaten bis hohen Einzelwerten („SkinBruise“ z.B. 0,69).
Diese Entwicklung zeigt deutlich, welchen Einfluss eine inhaltliche und methodische Abstimmung der BewerterInnen vor Beginn der Annotation auf die Qualität der Ergebnisse hat.

Daraus lässt sich schließen, wie essenziell ein Beobachtertraining für die Erstellung einer verlässlichen Ground Truth ist und wie gezielte Kommunikation zu einer Erhöhung der Übereinstimmung beiträgt.

Diese Ergebnisse lassen sich auch in der Fachliteratur wiederfinden. Eine Studie aus dem Jahr 2023 zeigt, wie sich die Interrater Reliability durch Trainingsmaßnahmen verbessern lässt. In diesem Fall beurteilen sechs BewerterInnen die methodische Qualität von bestimmten medizinischen Studien. Zuerst ohne spezifisches Training, dann nach gezielter Schulung mit konkreten Beispielen. Besonders bei schwierigen Fällen steigt danach die Übereinstimmungsrate deutlich an [15]. Die Autoren kommen hier zu dem Schluss, dass nicht nur gemeinsames Verständnis, sondern auch strukturiertes Training mit klaren Regeln und Beispielen die Interrater Reliability deutlich verbessern kann [15].

Um Beobachterabgleiche in Zukunft noch zuverlässiger zu gestalten, wäre es also nötig, eine vorangehende Trainingsphase mit Beispielannotationen durchzuführen. Außerdem können auch klar definierte Kriterien für die jeweiligen Labels und Beispielbilder hilfreich sein.

Insgesamt zeigt der Verlauf, dass ein Beobachterabgleich keine einmalige Maßnahme ist. Viel mehr kann dieser Vorgang als ein schrittweiser Prozess angesehen werden, bei dem ein systematischer fachlicher Austausch und gemeinsame Reflexion entscheidend für den Erfolg sind.

4.3 Schwierigkeiten und Optimierungsmöglichkeiten bei der Erstellung der Ground Truth

Während des Annotierungsprozesses innerhalb der ExpertInnengruppe treten im Verlauf der Entwicklung der Ground Truth mehrere methodische Herausforderungen auf, die eine wiederholte Anpassung in der Vorgehensweise erforderlich machen. Schon zu Beginn wird deutlich, dass der Ablauf der Erstellung der Ground Truth nicht optimal geplant ist.

Die gemeinsame Beurteilung der Bilder und das einheitliche Festlegen der Kriterien im Team hätten idealerweise vor dem eigentlichen Bewertungsprozess und dem Beobachterabgleich durch die beteiligten Diplomandinnen erfolgen sollen. So hätten potenzielle Unstimmigkeiten in der Bewertung vermieden und die Übereinstimmung in der Bildbewertung erhöht werden können. Außerdem hätte dieses Vorgehen für frühzeitige Sensibilisierung gesorgt, wie sie etwa bei der Unterscheidung von prozessbedingten und tatsächlich tierschutzrelevanten Läsionen von Nöten gewesen wäre.

In den ersten beiden Terminen zur Erstellung der *Ground Truth* werden sämtliche sichtbare Läsionen und Veränderungen auf den Schlachtkörpern überaus detailliert markiert und somit summiert sich bei einzelnen Bildern bis zu zehn Annotationen. Dabei werden auch viele

Hautläsionen markiert, welche vermutlich als prozessbedingt einzustufen sind. Das Ziel des Projektes - die Identifikation tierschutzrelevanter Läsionen - gerät dadurch in den Hintergrund. Dementsprechend wird intern im Team diskutiert, nur ausschließlich tierschutzrelevantes zu markieren, da dies dem übergeordneten Ziel des Projektes am ehesten entspricht. Infolgedessen werden die Markierungen auf wesentliche und vermeintlich tierschutzrelevante Veränderungen beschränkt.

Diese Vorgehensweise führt zwar zu einer deutlichen Reduktion der Markierungen aber somit auch einem fehlenden Standard in der Beurteilung und Befundung.

Eine zentrale Herausforderung besteht auch hier in der Unterscheidung zwischen Läsionen, die während der Lebenszeit des Tieres entstanden und solchen, die prozessbedingt durch die Schlachtung verursacht werden.

Die Unterscheidung zwischen frischen älteren, möglicherweise chronischen Läsionen, gestaltet sich aufgrund unterschiedlicher Aspekte als ebenso problematisch.

Sowohl die Farbe der Läsion, als auch die Tiefe sind häufig bei der bildbasierten Bewertung schwierig zu erkennen. Frische Läsionen lassen sich typischerweise durch ihre rötliche Färbung und ihr Auftreten an exponierten Körperregionen wie Schulter oder Flanke charakterisieren [16]. Diese Art der Verletzungen entsteht häufig durch Rangkämpfe oder Stress beim Transport oder dem Verladen der Tiere. Im Gegensatz dazu weisen ältere Hautläsionen häufig eine bräunliche oder gelbliche Verfärbung auf und sind eher an Ohren oder dem Rücken der Tiere lokalisiert. Diese Verletzungen können auf Verhaltensstörungen im Stall oder Mängel in der Haltung hinweisen [16].

Ein zusätzlicher Aspekt sind subjektive Unterschiede und natürliche Schwankungen innerhalb der Expertengruppe bei der Bewertung der verschiedenen Läsionen.

Die subjektive Beurteilung biologischer Merkmale ist selbst bei erfahrenen ExpertInnen variabel [17].

Hinzu kommen auch noch Faktoren wie die individuelle Tagesverfassung, unterschiedliche Lichtverhältnisse oder persönliche Erfahrungswerte.

Diese Erfahrungen decken sich auch mit einer aktuellen Studie aus dem Jahr 2023, in der elf IntensivmedizinerInnen unabhängig voneinander klinische Situationen bewerten sollen [18]. Auch in diesem Fall zeigen sich deutliche Unterschiede in der Bewertung und Cohens Kappa lag zwischen κ =0,25-0,38, was einer niedrigen Beobachter-Übereinstimmung entspricht [18]. Diese Studie kommt außerdem zu dem Schluss, dass es zur Erstellung einer verlässlichen *Ground Truth* nicht ausreicht, bei Uneinigkeit immer die Meinung der Mehrheit zu nehmen.

Stattdessen wird empfohlen, herauszufinden, welche Annotationen überhaupt zuverlässig genug sind, um daraus einen Algorithmus zu erstellen.
Laut den Autoren sind Labels dann zur Erstellung der *Ground Truth* besonders geeignet, wenn sie eine ausreichende Übereinstimmung zwischen verschiedenen ExpertInnen aufweisen, klar definiert sind und sich die Annotationen als wiederhol- und begründbar erweisen.
Bei zukünftigen Projekten könnte es also hilfreich sein, zuerst eine Voranalyse der Annotationen durchzuführen, um unklare Fälle zu identifizieren und gesondert zu behandeln.
Eine Möglichkeit könnte es sein, zunächst nur eine kleine Stichprobe der zu bewertenden Daten zu annotieren um daraus Cohen's Kappa zu ermitteln. Ergibt sich eine geringe Übereinstimmung können bei Bedarf die Kriterien nachträglich angepasst oder die BewerterInnen zusätzlich geschult werden. So wäre es möglich, Uneinigkeiten früher zu erkennen und somit eine stabile Grundlage zur Etablierung des Algoritmus zu schaffen.
Aufgrund der genannten Herausforderungen wird außerdem aktuell in Zusammenarbeit mit der CLK GmbH nach einer alternativen Methode gesucht um Hautläsionen zu markieren.
Anstatt Läsionen konkret zu annotieren wird vorgeschlagen, Bereiche mit einer Farbabweichung im Vergleich zur vermeintlich gesunden Haut des Tieres zu identifizieren und somit zukünftig eine einheitliche Bewertung zu ermöglichen.

4.4 Schlussfolgerung

Die vorliegende Arbeit zeigt, dass sowohl manuelle als auch bildbasierte Erhebungen von Hautläsionen bei Mastschweinen wertvolle Informationen liefern, die Ergebnisse jedoch in einzelnen Kategorien unterschiedlich stark übereinstimmen. Insbesondere die Beobachterabgleiche verdeutlichen, dass gezieltes Training und klare Definitionen der Läsionskategorien die Vergleichbarkeit und Zuverlässigkeit der Resultate deutlich verbessern können. Die Entwicklung einer *Ground Truth* gilt als Voraussetzung zur Etablierung eines funktionierenden Algorithmus, bringt aber unterschiedliche methodische Herausforderungen mit sich. Insgesamt lässt sich ableiten, dass bildbasierte Systeme und KI eine sinnvolle Ergänzung zur manuellen Befundung am Schlachthof darstellen, jedoch weitere Optimierungen notwendig sind, um eine konsistente und objektive Bewertung sicher stellen zu können.

5. Anhang

5.1 Literaturverzeichnis

1. Blömke L, Volkmann N, Kemper N. Evaluation of an automated assessment system for ear and tail lesions as animal welfare indicators in pigs at slaughter. Meat Sci. 2020;159:107934

2. D'Angelo M, Sciota D, Romano A, Rosamilia A, Guarnieri C, Cecchini C, Olivastri A, Marruchella G. Detecting ear lesions in slaughtered pigs through open-source convolutional neural networks. Porc Health Manag. 2025 May 23;11(1):29

3. aWISH Project. Animal welfare indicators at the slaughterhouse [Internet]. [Zugriff am 20. Februar 2025]. Verfügbar unter: https://www.awish-project.eu

4. Agrarmarkt Austria. Schweine – Nutztierhaltung & Haltungsformen [Internet]. AMA; [Zugriff am 19. Februar 2025]. Verfügbar unter: https://amainfo.at/tiere/nutztierhaltung-haltungsformen/schweine

5.CVAT [Internet]. Verfügbar unter: https://www.cvat.ai [Zugriff am 8. Juni 2025]

6. Republik Österreich. Tierkennzeichnungs- und Registrierungsverordnung 2009 (TKZVO 2009), BGBl. II Nr. 291/2009, § 9 Abs. 5 und 7. Zuletzt geändert durch BGBl. II Nr. 35/2011. Elektronische Fassung. Abgerufen am 31.12.2023 über RIS (Rechtsinformationssystem)

7. Goodfellow I, Bengio Y, Courville A. Deep learning. Cambridge (MA): MIT Press; 2016

8. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb). 2012;22(3):276-82. PMCID: PMC3900052

9. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas. 1960;20:37-46

10. Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. Information Processing & Management. 2009 Jul;45(4):427-37

11. Republik Österreich. Tierschutzgesetz – TSchG, BGBl. I Nr. 118/2004, zuletzt geändert durch BGBl. I Nr. 24/2023

12. Driessen B, Van Beirendonck S, Buyse J. The impact of grouping on skin lesions and meat quality of pig carcasses. Animals. 2020;10(4):544

13. Carroll GA, Boyle LA, Teixeira DL, van Staaveren N, Hanlon A, O'Connell NE. Effects of scalding and dehairing of pig carcasses at abattoirs on the visibility of welfare-related lesions. Animal. 2016;10(3):460-7

14. Welfare Quality. Welfare Quality assessment protocol for pigs [Internet]. Lelystad: Welfare Quality Network; 2009 [Zugriff am 8. Juni 2025]. Verfügbar unter: https://www.welfarequalitynetwork.net/media/1018/pig_protocol.pdf

15. Kaiser M, Kuijpers CCHJ, Fransen MP, van Doorn R. Improving the reliability of risk of bias assessments in prognostic studies using PROBAST: Results of a controlled experiment. BMJ Open. 2023 May;13(5):e070802

16. De Luca A, Magrin L, Scollo A. Abattoir-Based Measures to Assess Swine Welfare: Analysis of Consistency over Time and External Factors Influences. Animals. 2021;11(1):226

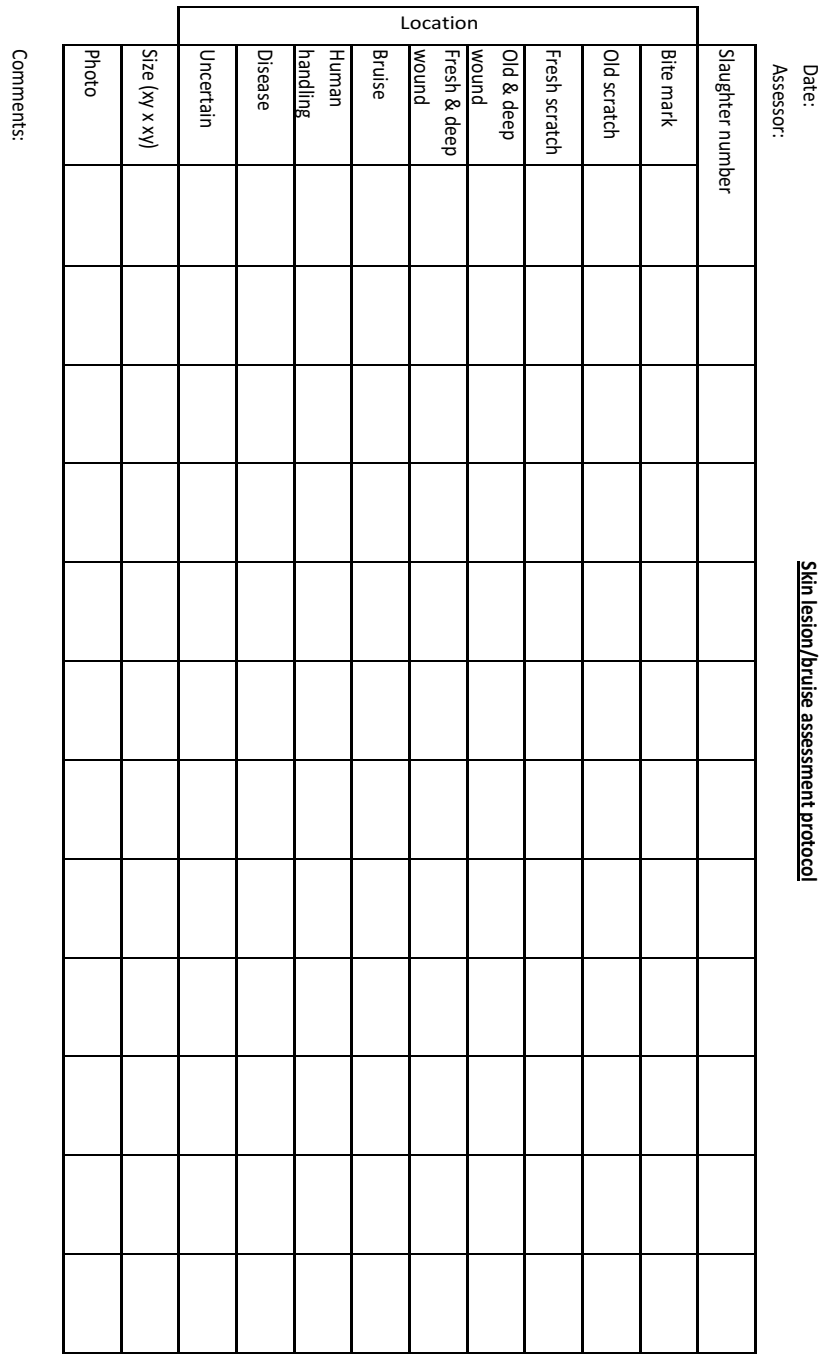
17. Sylolypavan A, Sleeman D, Wu H, Sim M. The impact of inconsistent human annotations on AI driven clinical decision making. NPJ Digit Med. 2023 Feb 21;6(1):26

18. Bushuven S, Bentele M, Gerber B, Michalsen A, Ilkilic I, Inthorn J. Single-rater reliability of a three-dimensional instrument for decision-making in tertiary triage and ICU-prioritization—a case vignette simulation study. BMC Anesthesiol. 2023 Jun 20;23(1):215

5.2 Abbildungsverzeichnis

Abbildung 1: Kamerasystem „PigInspector“ der Firma CLK GmbH am Standort Schwertberg 8
Abbildung 2: dorsale Ansicht der Schweineschlachtkörper im CVAT System 9
Abbildung 3: Skin Fresh Scratch 11
Abbildung 4: Skin Bruise 11
Abbildung 5: Skin Process Caused 11
Abbildung 6: Skin Rash 11
Abbildung 7: Skin Uncertain 12
Abbildung 8: Skin Old Scratch.....12
Abbildung 9: Tail Color Change Tip Without Tissue Loss 13
Abbildung 10: Tail Deep Wound.....13
Abbildung 11: Tail Bite Mark.....13
Abbildung 12: Tail Bruise.....13
Abbildung 13: Tail Superficial Wound 14
Abbildung 14: Tail Process Caused 14
Abbildung 15: Schematische Darstellung der Lokalisationen am Schlachtkörper 21
Abbildung 16 (links): Tierschutzrelevante Läsionen (Kratzer) an der Flanke eines Schweineschlachtkörpers.....29
Abbildung 17 (rechts): Tierschutzrelevante Läsion vermutlich durch unsachgemäße Anwendung des Elektroschockgerätes .....**Fehler! Textmarke nicht definiert.**
Abbildung 18 (links): prozessbedingte Läsionen am Rücken eines Tieres 30
Abbildung 19 (rechts): prozessbedingte Läsionen am Rücken eines Tieres 30

Tabelle 11: Protokoll zur Aufzeichnung der Hautläsionen



5.3 Tabellenverzeichnis

Tabelle 1: Definition der Hautläsionen (zur Verfügung gestellt von Dr. med. vet. Kristina Maschat) 6
Tabelle 2*:* Labels und Definition der Hautläsionen zur Bildannotierung in CVAT 9
Tabelle 3: Labels und Definition der Schwanzläsionen zur Bildannotierung in CVAT 12
Tabelle 4: Formeln und Definitionen der Kennzahlen zur Erstellung der Konfusionsmatrix [10] 17
Tabelle 5: Absolute Häufigkeiten und Prävalenzen der Hautläsionen der untersuchten Schlachtschweine (n=763) 20
Tabelle 6: Relative Häufigkeiten [%] der Lokalisation von Läsionen im Vergleich von Bildbewertung (CIVAT) und direkter Bewertung am Schlachtband 22
Tabelle 7: Absolute Häufigkeiten der Läsionstypen je Erfassungsmethode 22
Tabelle 8: Läsionen annotiert als tierschutzrelevant oder prozessbedingt 22
Tabelle 9: Ergebnisse des Beobachterabgleiches vom 23.07.2024 24
Tabelle 10: Ergebnisse des Beobachterabgleichs vom 30.07.2024 25
Tabelle 11: Protokoll zur Aufzeichnung der Hautläsionen 39