Sachbericht zum Verwendungsnachweis

I. Kurze Darstellung (max. 2 Seiten, wird veröffentlicht)

Zuwendungsempfänger /Ausführende Stelle:	Förderkennzeichen:	Laufzeit des Vorhabens:
Universität Leipzig / Professur Neuromorphe Informationsverarbeitung	28-D-K1-15A-20	10.11.2022– 19.05.2024

Verbundprojekt: Künstliche Intelligenz als Basis für die auf "Big Data" basierte Pflanzenzüchtung der Zukunft (KI-Zucht) - Teilprojekt D

1. Ursprüngliche Aufgabenstellung sowie der wissenschaftliche und technische Stand anden angeknüpft wurde.

In der Rapszüchtung ist es zunehmend notwendig, die Analyse komplexer „Big Data“ Anwendungen mit der Modellierung durch maschinelles Lernen (ML) zu koppeln, da Defizite bei der Vorhersage des Phänotyps zu einer geringen Züchtungsleistung, langen Wachstumsperioden und niedrigem Saatgutertrag und -qualität führen können. Dieses Problem hat seinen Ursprung darin, dass die klassischen statistischen Methoden, die von unseren bisher verwendet werden, wie z.B. multivariate und genomische Best Linear Unbiased Prediction (BLUP), eine geringe Effizienz bei der Modellierung und Vorhersage der nicht-deterministischen und nicht-linearen Natur der Interaktion zwischen dem Genotyp und der Umwelt haben, welche die Pflanzenphänotypen von Interesse erzeugen. Um diesen Herausforderungen zu begegnen, bestand die Hauptaufgabe unseres Teilprojekts darin, Modellierungsmethoden basierend auf ML zu untersuchen, um die Vorhersage für Raps-Phänotypen basierend auf genetischen Daten sowie Umweltdaten, die von unseren *KI-Zucht*-Partnern erzeugt und zusammengestellt wurden, zu verbessern. Die meisten der im Arbeitspaket 4 dieses Projekts weiterentwickelten, trainierten und validierten ML-Modelle wurden speziell für die Rapszüchtung angepasst.

2. Ablauf des Vorhabens.

Arbeitspaket 4 des Projekts war in die Meilensteine M18 und M19 aufgeteilt, die beide erfolgreich abgeschlossen wurden. Die Ziele von Meilenstein M18 waren der Entwurf und die Optimierung von ML-Algorithmen mit integrierten Datensätzen. Um dies zu erreichen, wurden vier Aufgaben entwickelt: (1) Vorverarbeitung von drei kanadischen Frühlingsraps-Datensätzen (Phänotyp-, Gen- und Umweltdaten), die von den Gruppen JLU-Gießen und NPZi zur Verfügung gestellt wurden, (2) Dimensionsreduktion und Integration der Datensätze, (3) (Teilweise manuelle) Merkmalsauswahl für das Training der ML-Modelle, (4) Entwurf und Optimierung verschiedener ML-Algorithmen und Künstlicher Neuronaler Netze (KNNs):

lineare/nicht-lineare, überwacht/nicht überwachte lernend, mit/ohne zeitliche Dimension. Die Ziele von Meilenstein M19 konzentrierten sich auf das Training und die Validierung unserer vielversprechendsten ML-Modelle, die ausgehend von Meilenstein 18 mit hochintegrierten Daten entwickelt wurden: Multilayer Perceptron (MLP), Convolutional Neural Networks (CNN) sowie Self-Organizing Maps (SOMs) für unüberwachtes Lernen. Es wurden vergleichende statistische Analysen für die Verfahren aus Meilenstein 18 durchgeführt, um die Leistung aller ML-Modelle, Schemata und die besten vorhergesagten Phänotypen aus Umweltdaten zu bewerten.

3. Wesentliche Ergebnisse sowie ggf. Zusammenarbeit mit anderen Stellen.

Bei der Analyse der Datensätze ohne Zeitdimension haben wir festgestellt, daß nichtlineare ML-Modelle bei der Vorhersage von Phänotypmerkmalen aus Umweltdaten deutlich besser abschneiden als lineare ML-Modelle. Von den neun Phänotypen, die aus den Jahr 2016 bis 2021 analysiert wurden, ist das Merkmal, das die Anzahl der Tage darstellt, die für die Reife der Schoten benötigt werden (d. h. „Reife“), mit allen von uns analysierten ML-Modellen durchweg der basierend auf Umweltdaten am besten vorhergesagte Phänotyp ist. Bemerkenswerterweise haben wir festgestellt, dass das überwachte MLP-Modell die niedrigsten Fehlerwerte für alle vorhergesagten Phänotypen aufweist und auch die höchste Vorhersagegenauigkeit für den „Samenertrag“ bietet, der durchweg der Phänotyp ist, der von allen anderen ML-Modellen am schlechtesten vorhergesagt wird. Bei der Analyse der Datensätze mit zeitlicher Dimension haben wir festgestellt, dass überwachte lernende CNNs durchweg für alle Phänotyp-Merkmale bessere Vorhersagen liefern als unüberwacht lernende SOMs. Weitere umfangreiche Analysen bestätigen, dass MLP und CNNs die besten Methoden zur Vorhersage von „Saatgutertrag“ und „Reife“ sowie anderen relevanten Merkmalen zur Verbesserung der Saatgutqualität sind, wie z. B. dem Gehalt an Glucosinolaten, Öl und Protein in den Samen. Zusätzlich konnte festgestellt werden, daß SOMs nicht nur ein nützliches Werkzeug zur visuellen Interpretation der integrierten Daten in den erzeugten SOM-Karten sind, sondern auch auf Korrelation zwischen Wetter und Ertrag hinweisen können.

Durch die Verwendung unserer wichtigsten Erkenntnisse und Protokolle zur Reproduktion unserer besten ML-Modelle werden Züchter und ML-Experten von *KI-Zucht-Partnern* in der Lage sein, effizienter und mit weniger erforderlichen Analysen vorherzusagen, welche Rapshybriden einen höheren Ertrag und eine höhere Saatgutqualität aufweisen könnten, die von bestimmten Umweltvariablen signifikant beeinflusst werden. Unsere Ergebnisse wurden während des Förderzeitraums mit Partnern aus den vom BMEL geförderten KI-Projekten und X-KIT diskutiert, und unsere gemeinsamen Ergebnisse werden demnächst in internationalen Fachzeitschriften veröffentlicht.

Sachbericht zum Verwendungsnachweis

II. Eingehende Darstellung (max. 20 Seiten, wird veröffentlicht)

Zuwendungsempfänger /Ausführende Stelle:	Förderkennzeichen:	Laufzeit des Vorhabens:
Universität Leipzig / Professur Neuromorphe Informationsverarbeitung	28-D-K1-15A-20	10.11.2022 – 19.05.2024

Verbundprojekt: Künstliche Intelligenz als Basis für die auf "Big Data" basierte Pflanzenzüchtung der Zukunft (KI-Zucht) - Teilprojekt D

1. Rahmen und Erzielte Ergebnisse des Vorhabens durchgeführten Arbeiten, insbesondere im Vergleich zur ursprünglichen Vorhabenbeschreibung.

AP4: Untersuchung (un-)überwachter Verfahren des maschinellem Lernens (ML) zur Erkennung wichtiger Eingabeparameter.

Arbeitspaket 4 des Projekts war in die Meilensteine M18 und M19 unterteilt, die erfolgreich abgeschlossen wurden. Die Ziele von Meilenstein M18 waren der Entwurf und die Optimierung von ML-Algorithmen mit integrierten Datensätzen. Die Ziele von Meilenstein M19 konzentrierten sich auf das Training und die Validierung unserer vielversprechendsten ML-Modelle, die von Meilenstein 18 mittels hochintegrierten Daten entwickelt wurden: Multilayer Perceptron (MLP), Convolutional Neural Networks (CNN) für überwacht lernende Modelle sowie Self-Organizing Maps (SOMs) für unüberwachtes Lernen. Es wurden vergleichende statistische Analysen für die Verfahren aus Meilenstein 18 durchgeführt, um die Leistung aller ML-Modelle, Schemata und die besten vorhergesagten Phänotypen aus Umweltdaten zu bewerten. Die in den Meilensteinen M18 und M19 durchgeführten Arbeiten werden im Folgenden in fünf Hauptaufgaben aufgeteilt:

(1) Dimensionalitätsreduktion der Datensätze. Zunächst wurde eine Vorverarbeitung der drei von der NPZi-Gruppe zur Verfügung gestellten Frühjahrsraps-Datensätze aus Kanada durchgeführt. Dieser Schritt umfasste die Imputation fehlender Daten, die Vorhersage allgemeiner statistischer Schätzer und die Integration der drei Datensätze (Phänotyp-, Umwelt- und SNP-Daten) für jedes Feld und jedes Jahr, das von 2016 bis 2021 beprobt wurde. Es wurde eine Matrix generiert, die alle Phänotyp-Merkmale mit den Umweltvariablen und IDs-Indizes abbildet, welche 6.415 Hybriden mit SNPs-Daten (Hybrid-SNPs-Index) von 2016 bis 2021 darstellt. Dieser Ansatz führte zu einer Reduktion um 66,6% der ursprünglichen Datensätze (Abb. 1a).

(2) Merkmalsauswahl für das Training von ML-Modellen mit überwachte Lernen OLS-Analysen. Die Merkmalsauswahl, die Erkennung von Kovarianzen sowie die zeitliche Abhängigkeit zwischen Umweltmerkmalen und Phänotypen wurden zunächst mit dem Pearson-Korrelationskoeffizienten und Ordinary Least Squares (OLS) Analysen geschätzt. Um Korrespondenzen mit dem Wachstumszyklus des Frühjahrsraps-Morphotyps zu finden (Abb. 1b), wurden paarweise OLS-Regressionen durchgeführt. Damit wurde der Anteil der Varianz geschätzt, der durch die vorhergesagten Zusammenhänge zwischen den Phänotypen und den einzelnen Umweltmerkmalen während des Jahres erklärt wird (Mittelwert für alle Jahre und Zeitreihen pro Monat). Wie aus Abb. 1c hervorgeht, können Temperatur, Wasserverfügbarkeit, Luftfeuchtigkeit sowie Wolkendichte die bei den wichtigsten Phänotypen beobachteten Schwankungen (30-60 %) am besten erklären, z. B. die Anzahl der Tage, die für die Reife der Schoten (Samenfüllung) und die Blüte benötigt werden, die Pflanzenhöhe, sowie Öl- und Proteingehalt der Samen. In guter Übereinstimmung mit dem Wachstumszyklus im Frühjahr, lassen sich die für den Reifephänotyp beobachteten Schwankungen am besten durch Veränderungen der mittleren Temperatur, der Luftfeuchtigkeit und der Wolkendichte von Juni bis August vorhersagen. Wie erwartet können der Ertrag und der Glucosinolatgehalt der Samen (die komplexesten Phänotypen) durch OLS nicht genau vorhergesagt werden.

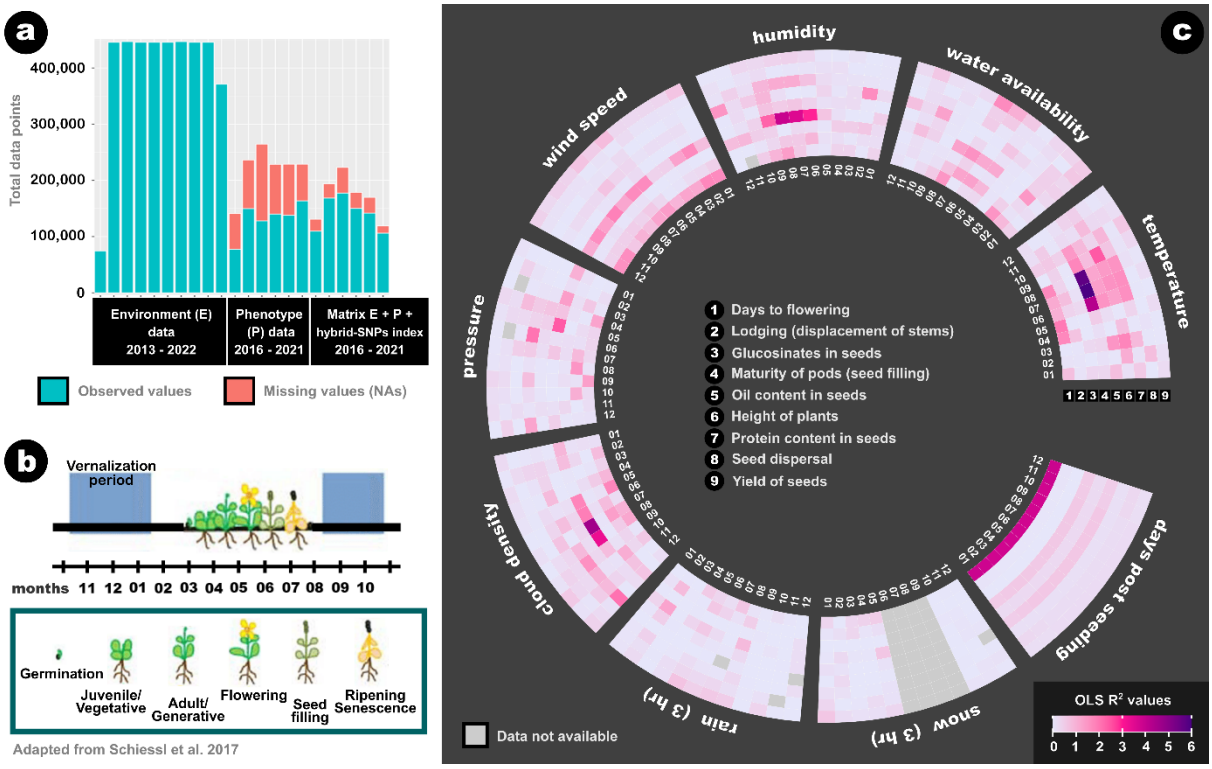

Abbildung 1. (a) Dimension der Datensätze. (c) Bestimmungskoeffizienten (R^2_{OLS}), die angeben, wie viel der Varianz des Phänotyps durch ausgewählte Umweltvariablen vorhergesagt werden kann, je nach Monat (01-12) und Wachstumszyklus des Frühjahrsraps (b).

(3) Implementierung, Training und Validierung von linearen und nichtlinearen ML-Modellen ohne Zeitdimension. Auf der Grundlage der in (1) und (2) beschriebenen Arbeit wurde der Datensatz von Umwelt- und Phänotypdaten zufällig in einen Lernsatz, einen Validierungssatz und einen Testsatz aufgeteilt. Sieben lineare und nichtlineare ML-Modelle ohne Zeitdimension wurden speziell für die Vorhersage von Raps-Phänotypen aus Umweltdaten als Regressionsaufgabe generiert. Darunter fallen die Lineare Regression mit additiver Verzerrung (LR), Bayessche Lineare Regression (BLR) sowie nicht-lineare ML-Modelle wie Random Forest (RF: 100 Schätzer), k-Nächster Nachbar (KNN), Support Vector Machine (SVM) mit einem Radial Basis Function Kernel (RBFk), Entscheidungsbaum (DT), und Multilayer Perceptron (MLP) mit Aktivierungsfunktion für gleichgerichtete lineare Einheiten (ReLU) und 2 versteckte Schicht (1000 Neuronen und 500 Neuronen). Für alle ML-Modelle wurde ein Merkmalsvektor erstellt, indem der Mittelwert jeder der neun Umweltvariablen über jedes Jahr und jeden Standort, dem ein Hybrid zugeordnet ist, zusammengefasst wurde. Eine „Min-Max-Merkmalsskalierung“ wurde verwendet, um die Daten im Bereich (0,1) zu normalisieren. Alle ML-Modelle wurden mit einem 80:20-Datensatzsplit trainiert und getestet. Die Vorhersagegenauigkeit und die Leistung jedes ML-Modells wurden anhand des mittleren quadratischen Fehlers (RMSE) und des Bestimmtheitsmaßes (R^2) zwischen vorhergesagten und tatsächlichen Werten bewertet. Anhand des RMSE-Scores können wir erkennen, welche ML-Modelle aufgrund von großen Fehlern (Ausreißern) stärker benachteiligt sind.

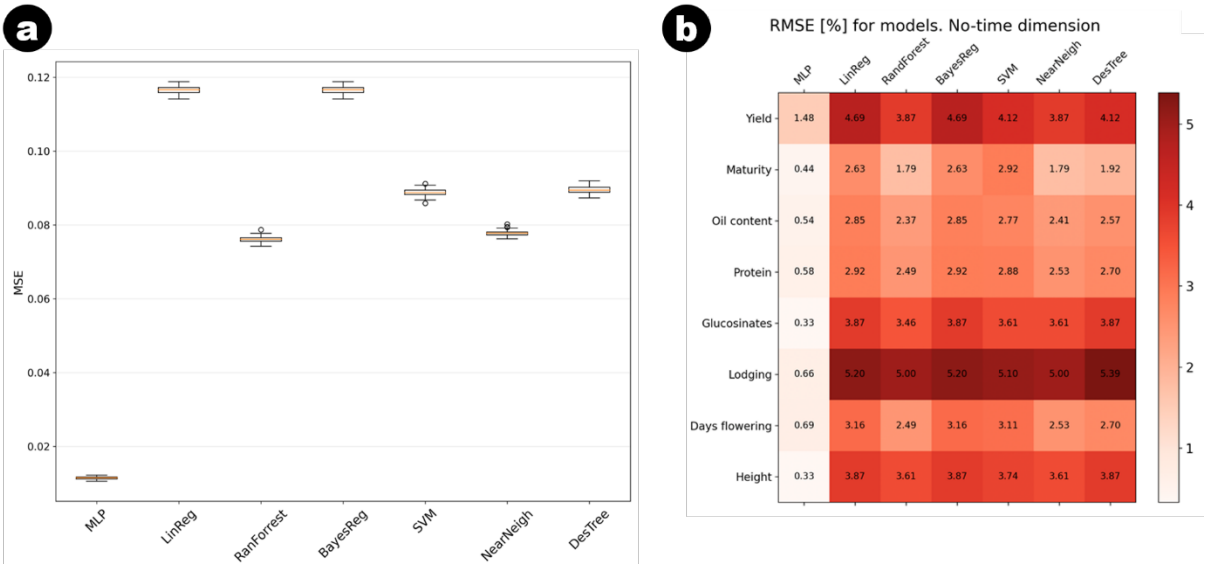

Abbildung 2. (a) Das Balkendiagramm zeigt den Vergleich der MSE-Werte, die die Leistung aller ML-Modelle (keine zeitliche Dimension) zur Vorhersage von Erträgen aus Umweltvariablen über alle Jahre hinweg bewerten. (b) Die Heatmap zeigt den Vergleich der RMSE-Werte (%), mit denen die Leistung aller ML-Modelle (keine zeitliche Dimension) zur Vorhersage von neun Rapsphänotypen anhand von Umweltvariablen über alle Jahre hinweg bewertet wird.

Unsere Ergebnisse zeigen, dass nichtlineare Modelle wie RF, SVM-RBFk, DT, KNN und MLP deutlich besser abschneiden als die linearen LR- und BLR-Modelle. Allerdings nähern sich die linearen Modelle dem Mittelwert der Testdaten am besten an, während die nichtlinearen Modelle die Varianz besser beschreiben (Abb. 2a). Von den neun Phänotypen, die in jedem Jahr analysiert wurden, ist „Reife“ mit allen ML-Modellen durchweg der am besten aus den Umweltdaten vorhergesagte Phänotyp, während der Saatgutertrag und die Lagerfähigkeit bei den meisten ML-Modellen die am schlechtesten vorhergesagten Phänotypen sind. Bemerkenswerterweise weist das MLP nicht nur die niedrigsten Fehlerwerte für alle aus Umweltdaten vorhergesagten Phänotypen auf, sondern liefert auch die höchste Vorhersagegenauigkeit für den „Saatgutertrag“, der als der komplexeste Phänotyp unseres Datensatzes gilt (Abb. 2b).

(4) Training von ML-Modellen mit unüberwachte Lernen Self-Organizing Maps (SOMs). SOMs (Abb. 3a) wurden trainiert, um Phänotypen mit Zeitreihen für jede Umweltvariable pro Jahr vorherzusagen, welche mit überlappenden Zeitfenstern (von einer Woche und Schritten von einem Tag) und der Aggregation von Mittelwerten für jedes Fenster erstellt wurden. Vier Schemata von Datensätzen wurden getestet und anhand des RMSE zwischen vorhergesagten und tatsächlichen Werten für Validierungsdatensätze bewertet (Abb. 3b-c). Wie in Abbildung 2c dargestellt, sind die Phänotypen der „Saatgutqualität“ (insbesondere der Glucosinolatgehalt) und des Lagerverhaltens (Stängelverschiebung) in allen Jahren und in allen untersuchten Schemata die am besten aus Umweltdaten vorhergesagten Phänotypen, während der “Saatgutertrag” durchweg der am schlechtesten vorhergesagte Phänotyp ist. Bemerkenswert ist, dass die vorhergesagten Phänotypen mit den niedrigsten Fehlerwerten für alle SOMs erhalten wurden, die mit dem Schema von Datensätzen trainiert wurden, die Hybrid-SNPs-Indizes und eine reduzierte Anzahl von Umweltmerkmalen haben. Zusätzlich konnte festgestellt werden, daß SOMs nicht nur ein nützliches Werkzeug zur visuellen Interpretation der integrierten Daten in den erzeugten SOM-Karten sind, sondern auch auf Korrelation zwischen Wetter und Ertrag hinweisen können.

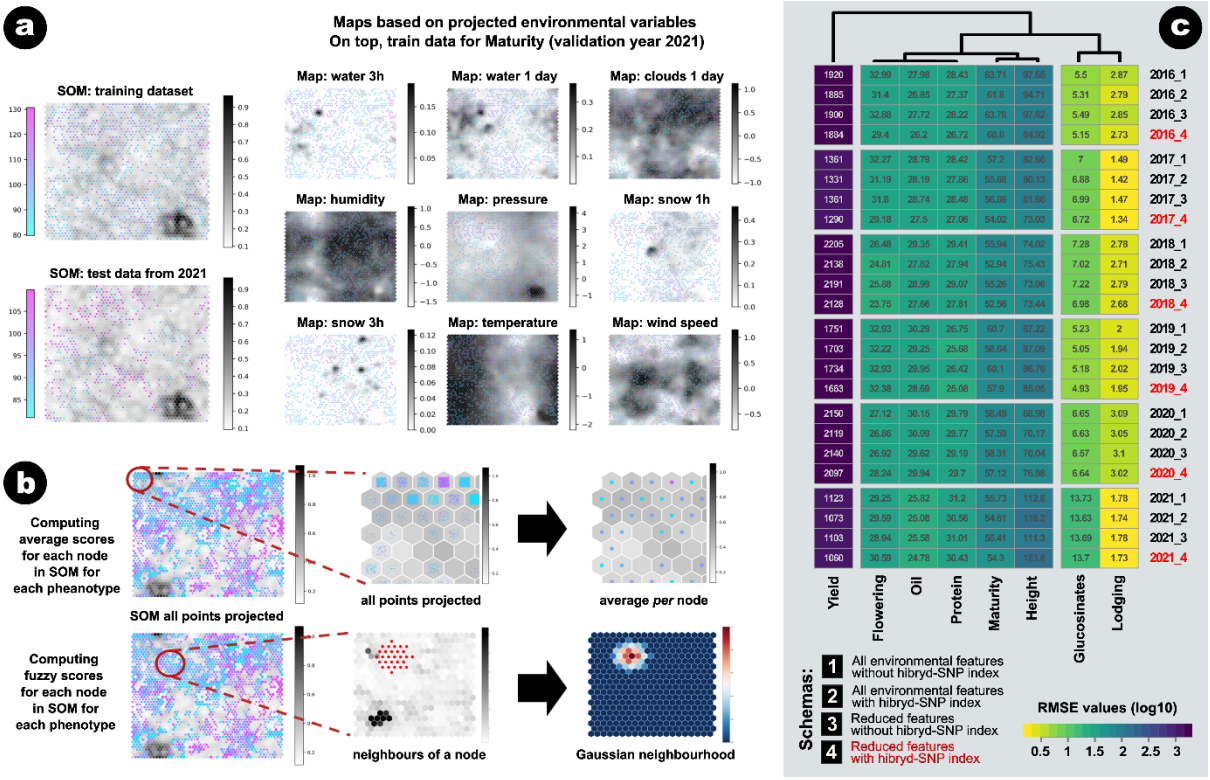

Abbildung 3. (a) Beispiel für SOMs (50x50 Neuronen), die mit ausgewählten Umweltmerkmalen trainiert wurden, wobei die Daten des Reifephänotyps farbig auf die Karten projiziert wurden. (b) Berechnung der RMSE-Werte unter Verwendung von Fuzzy-Scoring für Validierungsdatensätze. (c) RMSE-basierte Vorhersagegenauigkeit für 4 Schemata von Datensätzen.

(5) Performanz des Convolutional Neural Networks (CNN) unter Berücksichtigung der Zeitdimension. CNN (Abb. 3a) wurden trainiert, um Phänotypen mit Zeitreihen für jede Umweltvariable pro Jahr vorherzusagen, welche ebenfalls mit dem überlappenden Zeitfenstern (von einer Woche und Schritten von einem Tag) und der Aggregation von Mittelwerten für jedes Fenster erstellt wurden (Abb. 3a). Drei Architekturmodelle von Datensätzen wurden getestet und anhand des RMSE zwischen vorhergesagten und tatsächlichen Werten für Validierungsdatensätze bewertet. Unsere Ergebnisse zeigen, dass die Anzahl der Tage, die für die „Schotenreife“ (Samenfüllung) und die Blüte benötigt werden, sowie der Gehalt an Glucosinolaten, Öl und Protein in den Samen, die am besten vorhergesagten Phänotypen mit CNNs aus Umweltvariablen über alle analysierten Jahre und Arquitecture Modelle sind (Abb. 3b). Im Gegensatz dazu war das Merkmal, welches die permanente Verschiebung eines Stängels („lodging“) darstellt sowie die Höhe der Pflanze die am schlechtesten vorhergesagten Phänotypen mit dem CNN-Modell aus Umweltdaten (Abb. 3b). Beim Vergleich der Leistung von CNNs und SOMs mit zeitlicher Dimension haben wir festgestellt, dass CNNs durchweg bessere ML-Modellvorhersagen

liefern als unüberwacht lernende SOMs und dies für alle Phänotyp-Merkmale in allen untersuchten Jahren und Schemata (Abb. 3c).

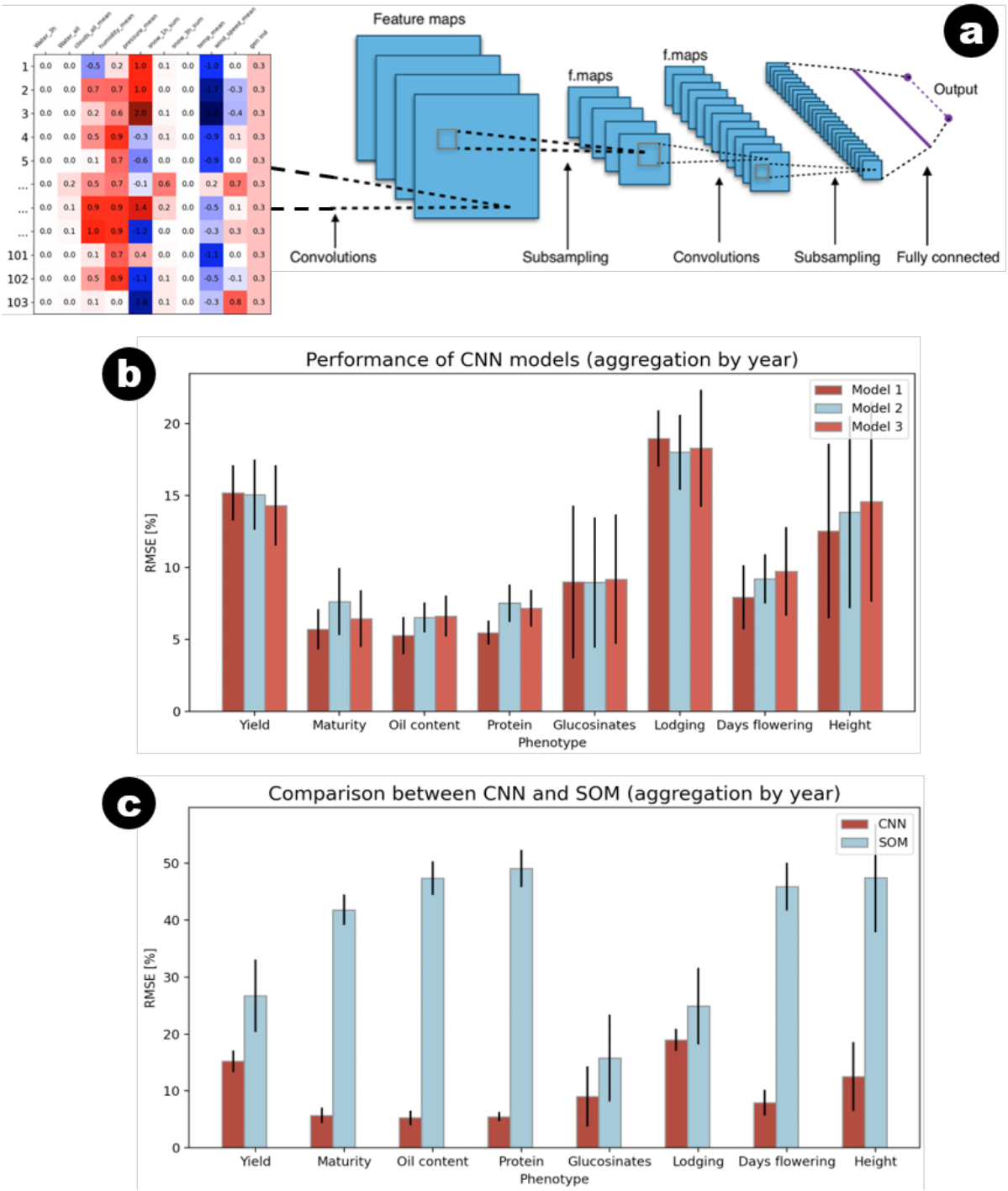

Abbildung 3. (a) Schematische Darstellung eines CNN-Modells mit dem Phänotyp und den Umweltdaten als Eingabe. (b) Der Boxplot zeigt die Leistung der Vorhersage der Phänotypen mit dem CNN-Modell und drei Architekturen. (c) Der Boxplot zeigt den Vergleich der Leistung der CNN- und SOM-Modelle bei der Vorhersage der Phänotypen aus Umweltdaten.

Schlussfolgerungen

Es wurde festgestellt, dass mittels überwachter nichtlinearer ML-Modelle - mit oder ohne Berücksichtigung der zeitlichen Dimension - die Vorhersage von Raps-Phänotypen mit nicht-deterministischer Natur, die von mehreren Umwelt- und genetischen Variablen erheblich beeinflusst werden, tatsächlich um mehr als eine Größenordnung verbessert werden kann. Die Ergebnisse bestätigen darüberhinaus, dass der „Samenertrag“ der komplexeste quantitative Phänotyp ist, welcher mit einem ML-Modell ebenfalls vorhergesagt werden kann. Bemerkenswerterweise haben wir festgestellt, dass die überwachten ML-Modelle, das MLP und das CNN, die besten Prädiktoren für den „Saatgutertrag“ aus Umweltdaten sind. Der „Saatgutertrag“ ist jedoch für die Züchter nur bedingt vererbbar und direkt selektierbar, da er in der Regel von mehreren Genen gesteuert wird und auch stark von der Umwelt beeinflusst wird (Niazian & Niedbała, 2020). Daher war es von Bedeutung, dass unsere MLP- und CNN-Modelle auch in der Lage waren, die Vorhersage anderer Merkmale, die mit dem „Saatgutertrag“ und der „Saatgutqualität“ zusammenhängen, erheblich zu verbessern, wie etwa die Anzahl der Tage, die für die „Schotenreife“ (Samenfüllung) und die Blüte benötigt werden, sowie der Gehalt an Glucosinolaten, Öl und Protein in den Samen. Anstatt sich ausschließlich auf den Phänotyp „Saatgutertrag“ zu konzentrieren, können Pflanzenzüchter diese „einfacheren“, höher korrelierten Phänotypen vorhersagen und nutzen, um einen größeren Einfluss auf den „Samenertrag“ und die „Samenqualität“ mit Nährwert für die menschliche Gesundheit haben.

Die in diesem Bericht vorgestellten Ergebnisse stehen in guter Übereinstimmung mit neueren Arbeiten, die zeigen, dass das MLP-Modell die Vorhersage des „Samenertrags“ und anderer verschiedener Phänotypen bei Raps tatsächlich ein RMSE-Fehler von 7,5 % erreicht werden kann (Niedbała, 2019). Andererseits wurde das CNN-Modell erst kürzlich bei einigen anderen Pflanzen zur automatischen Erkennung von Pflanzenkrankheiten (Wang et al. 2021, Xu et al. 2022, Zhou et al. 2022) und zur Pflanzenauswahl in Züchtungsprogrammen (Liu et al., 2019, Wang and Su 2022, Xie et al., 2024) untersucht. Unsere Arbeit zeigt zum ersten Mal, dass die CNN-Modelle eine höhere Vorhersagegenauigkeit bei der Vorhersage der komplexesten Raps-Phänotypen haben als SOMs und andere ML-Modelle. Beispielsweise waren die Nichtlinearität der Daten, die Abhängigkeit von der Zeitdimension und die Komplexität des vorhergesagten Phänotyps Grund dafür, dass die Implementierung der für den Meilenstein M18 vorgeschlagenen ART Fuzzy Networks erfolglos war. Im Vergleich zu anderen ML-Modellen besteht ein weiterer Vorteil der Verwendung von CNN darin, dass das

Problem der erforderlichen großen, manuell erstellten Merkmale durch die Verwendung von Bilderweiterungen und kleinen, manuell kommentierten empirischen Datensätzen für die Feinabstimmung eines synthetischen Bootstrapped CNN (Barth et al., 2019) gelöst werden kann.

Durch die Verwendung unserer wichtigsten Erkenntnisse und Protokolle zur Reproduktion unserer besten ML-Modelle werden Züchter und andere ML-Experten der Projektpartner KI-Zucht in der Lage sein, effizienter und mit weniger Analyseaufwand vorherzusagen, welche Rapshybriden mit potenziell höherem Ertrag und besserer Saatgutqualität stärker von bestimmten Umwelt- und genetischen Variablen beeinflusst werden. Insgesamt zeigen unsere Ergebnisse, dass der Einsatz von nicht-linearen überwachten Deep-Learning-Modellen im Vergleich zu traditionellen Methoden vielversprechende Ergebnisse für die Raps-Phänotypen-Selektion in Zuchtprogrammen liefert.

Zitierte Literatur

Barth, R., IJsselmuiden, J., Hemming, J. and Van Henten, E.J., 2019. Synthetic bootstrapping of convolutional neural networks for semantic plant part segmentation. Computers and Electronics in Agriculture, 161, pp.291-304.

Liu, Y., Wang, D., He, F., Wang, J., Joshi, T. and Xu, D., 2019. Phenotype prediction and genome-wide association study using deep convolutional neural network of soybean. Frontiers in genetics, 10, p.1091.

Niazian, M. and Niedbała, G., 2020. Machine learning for plant breeding and biotechnology. Agriculture, 10(10), p.436.

Niedbała, G., 2019. Application of artificial neural networks for multi-criteria yield prediction of winter rapeseed. Sustainability, 11(2), p.533.

Wang, P., Niu, T., Mao, Y., Zhang, Z., Liu, B. and He, D., 2021. Identification of apple leaf diseases by improved deep convolutional neural networks with an attention mechanism. Frontiers in Plant Science, 12, p.723294.

Wang, Y.H. and Su, W.H., 2022. Convolutional neural networks in computer vision for grain crop phenotyping: A review. Agronomy, 12(11), p.2659.

Xie, Z., Weng, L., He, J., Feng, X., Xu, X., Ma, Y., Bai, P. and Kong, Q., 2024. PNNGS, a multi-convolutional parallel neural network for genomic selection. Frontiers in Plant Science, 15, p.1410596.

Xu, C., Wang, X. and Zhang, S., 2022. Dilated convolution capsule network for apple leaf disease identification. Frontiers in Plant Science, 13, p.1002312.

Zhou, H., Deng, J., Cai, D., Lv, X. and Wu, B.M., 2022. Effects of image dataset configuration on the accuracy of rice disease recognition based on convolution neural network. Frontiers in Plant Science, 13, p.910878.

2. Wichtigste Positionen des zahlenmäßigen Nachweises.

Die finanzielle Unterstützung des BMEL/BLE zur Durchführung der Forschungsarbeiten von Meilenstein M18 und M19 aus Arbeitspaket 4 des Projektes KI-Zucht wurde für zwei Arten von Ausgaben verwendet: (1) die Einstellung von wissenschaftlichem und technischem Personal zur Durchführung der in Abschnitt II.1 dieses Berichts beschriebenen Forschungsarbeiten (siehe auch die beigefügten Dateien für den zahlenmäßigen Nachweis dieses Berichts) und (2) die Anschaffung einer Hochleistungs-Workstation.

Für die Erstellung der ML-Modelle war es zwingend nötig, einen entsprechenden Hochleistungsrechner zu beschaffen, damit das unter (1) eingestellte Personal insbesondere unter Berücksichtigung des verspäteten Eintritts der Arbeitsgruppe in das bereits laufende Projekt so effizient wie möglich arbeiten kann. Für das Training und die Validierung unserer ML-Modelle sowie für die Analyse großer Mengen komplexer und nicht-deterministischer Informationen mit „mehreren unabhängigen Variablen gegenüber mehreren abhängigen Variablen“ ist ein Hochleistungsrechner erforderlich. Die erforderliche Rechenzeit und -leistung, um die Forschungsarbeiten von Meilenstein M18 und Meilenstein M19 fristgerecht abzuschließen und die (mögliche) Weiterarbeit an diesem Projekt fortzusetzen, war und ist nur mit einer solchen adäquaten Rechnerkapazitäten sowie dem entsprechend erfahrenen Personal möglich. Zu diesem Zweck wurde ein Team von drei hochqualifizierten Informatikern und einigen zusätzlichen Technikern aus den Bereichen maschinelles Lernen, Bioinformatik und Pflanzengenomik für den Zeitraum von Februar 2023 bis Mai 2024 eingestellt, da sich die

Einstellung aufgrund der späten Bewilligung des Teilprojekts Ende November 2022 administrativ verzögert hatte (siehe Jahresbericht für 2022). Daher hat unser interdisziplinäres Forschungsteam die gesamte Forschungsarbeit für die Meilensteine M18 und M19 (wie in Abschnitt II.1 dieses Berichts beschrieben) effektiv in 14 Monaten durchgeführt, anstatt der ursprünglich geplanten 24 Monate. Entsprechend musste ein erhöhter Personalaufwand mit höheren Kosten zur Bewältigung der gleichen Arbeitsziele in der stark verkürzten Zeit eingeführt werden.

Es ist wichtig zu erwähnen, dass das interdisziplinäre Forschungsteam, das für dieses Projekt gebildet wurde, entscheidend für die Entwicklung des besten biologischen und rechnerischen Ansatzes war, um die große Menge an Eingabedaten, die von unseren KI-Zucht-Partnern zur Verfügung gestellt wurden (abiotische, genetische und phänotypische Daten), rechtzeitig und adäquat erfolgreich zu bearbeiten. Das nächste Ziel des interdisziplinären Forschungsteams bestand darin, die extreme Datenmenge in unsere ML-Modelle zu integrieren und die besten manuell identifizierten Merkmale für das Training und die Validierung unserer komplexen ML-Methoden zu analysieren und auszuwählen. Der hohe Grad an Neuartigkeit und Innovation der in den Meilenstein M18 und M19 erzielten Ergebnisse – in der Hälfte der ursprünglich geplanten Zeit – war nur durch die externe Finanzierung des BMEL/BLE möglich, da die Abteilung Neuromorphe Informationsverarbeitung und die Universität Leipzig nicht über ausreichende personelle, finanzielle und rechnerische Kapazitäten verfügen, um dieses Projekt alleine zu finanzieren und damit durchzuführen.

3. Notwendigkeit und Angemessenheit der geleisteten Projektarbeit.

Die Motivation, die Forschungsarbeiten des Arbeitspakets 4 in das KI-Zucht-Projekt einzubinden, entstand aus der Notwendigkeit, die Analyse großer und komplexer Datenmengen mit Hilfe von ML zu verarbeiten, um die Züchtung von Raps mit automatisierten Methoden voranzubringen. Die klassischen statistischen Methoden, die von unseren KI-Zucht-Partnern üblicherweise verwendet werden, wie z. B. multivariate und genomische Best Linear Unbiased Prediction (BLUP), erreichen eine geringere Effizienz bei der Modellierung der nichtlinearen Natur komplexer Daten und eine geringere Vorhersagegenauigkeit von Phänotypen mit Feldfrucht Interessen, was in der Regel im Endeffekt zu geringerer Zuchteffizienz, langen Wachstumsperioden und geringerem Saatgutertrag. Wir haben diese Herausforderungen über die Meilensteine M18 und M19 hinweg untersucht und erfolgreich bewältigt, indem wir zunächst die Dimensionalität der komplexen Datensätze, die von unseren KI-Zucht-Partnern bereitgestellt wurden, in Matrizen reduziert haben, welche die phänotypischen, genetischen und umweltbezogenen Informationen der in Kanada beprobten Rapshybriden effizient integrieren. Anschließend haben wir verschiedene lineare und

nichtlineare ML-Modelle, insbesondere neuartige Deep-Learning-Modelle, weiterentwickelt, trainiert und validiert, um eine höhere Genauigkeit bei der Vorhersage von Raps-Phänotypen zu erreichen, die von den genetischen Daten sowie Umweltdaten aufgezeichneten Veränderungen erheblich beeinflusst werden. Wir konzentrierten uns bei unserer Forschungsarbeit auf komplexe ML-Modelle, die von den weiteren Mitgliedern des KI-Zucht-Konsortiums noch nicht untersucht wurden und die erst kürzlich von einigen anderen internationalen Forschungsgruppen in die Pflanzenzüchtung integriert wurden. Wie in den meisten Rapszüchtungsprogrammen sind „Saatgutertrag“ und „Saatgutqualität“ die bei weitem am komplexesten vorherzusagenden Phänotypen und waren daher die Hauptphänotypenziele unserer ML-Modellierung.

Wie in Abschnitt II.1 dieses Berichts beschrieben, haben wir herausgefunden, dass überwacht lernende nichtlineare ML-Modelle - entweder mit oder ohne Berücksichtigung der zeitlichen Dimension - die Vorhersage von Raps-Phänotypen mit nicht-deterministischer Natur, die von mehreren Umwelt- und genetischen Variablen erheblich beeinflusst werden, tatsächlich um mehr als eine Größenordnung verbessern können. Bemerkenswerterweise haben wir festgestellt, dass die überwachten ML-Modelle, das mehrschichtige Perzeptron (MLP) und das CNN, die besten Prädiktoren für den „Saatgutertrag“ und die „Saatgutqualität“ aus Umweltdaten sind. Ebenfalls wurden die Anzahl der Tage, die für die „Schotenreife“ (Samenfüllung) und die Blüte benötigt werden, sowie der Gehalt an Glucosinolaten, Öl und Protein in den Samen am besten durch MLP- und CNN-Modelle vorhergesagt. Da alle diese vorhergesagten Phänotypen für die Verbesserung von Kulturpflanzen und die Lebensmittelindustrie von entscheidender Bedeutung sind, bieten unsere Ergebnisse den Züchtern und anderen ML-Experten ein großes Potenzial, nicht nur die Merkmalsauswahl für die Entwicklung neuer Rapssorten mit höherem Nährwert für die menschliche Gesundheit, höherer Ertragsleistung und größerer Widerstandsfähigkeit gegenüber dem Klimawandel zu fördern, sondern auch ein automatisierte Vorhersage mit Hilfe von ML-Verfahren zu ermöglichen (siehe auch Abschnitt II.4 dieses Berichts).

4. Verwertung

Es wurden Protokollberichte mit umfassenden Informationen erstellt, um unsere besten trainierten und validierten ML-Modelle mit integrierten und ausgewählten Eingabedaten aus den Originaldatensätzen für unsere Partner JLU-Gießen und NPZi reproduzierbar für zukünftige Anwendungen der Partner zu machen. Da keine essentiellen neuen Programmierteile erforderlich waren, um die maschinelle Lernmodelle für Meilensteine M18 und M19 anzupassen, enthalten die Protokollberichte, die unseren NPZi-Partnern zur Verfügung gestellt wurden, klare Anweisungen und Leitlinien für die Erstellung der

Architekturen unserer ML-Modelle und die Einstellungen unserer besten identifizierten Architekturen mit offen zugänglichen Online-Quellen (d.h. z.B. TensorFlow, Keras, PyTorch etc.). Unsere Protokolle enthalten auch zusätzliche Informationen über die Vorverarbeitung und das Training der von unseren JLU-Gießen- und NPZi-Partnern generierten Datensätze sowie die Mindestrichtlinien, die für die Interpretation der aus unseren ML-Modellen abgeleiteten Ergebnisse erforderlich sind.

Durch die Verwendung unserer wichtigsten Erkenntnisse und Protokolle zur Reproduktion unserer besten Deep-Learning-Modelle werden Züchter und andere ML-Experten von KI-Zucht in der Lage sein, mit höherer Genauigkeit und weniger zusätzlichen Analysen vorherzusagen, welche ihrer aktuellen Rapshybriden Phänotypen ausprägen könnten, die von bestimmten Umweltvariablen signifikant beeinflusst werden können und potenziell einen höheren „Saatgutertrag“ und/oder eine höhere „Saatgutqualität“ erzeugen. Die Verwendbarkeit und Effizienz unserer ML-Modelle und Protokolle können in der Zukunft von unseren KI-Zucht-Partnern weiterentwickelt werden. Unsere Zusammenarbeit mit den Projektpartnern soll auch dazu dienen, eine effizientere Integration genetischer Marker in unsere ML-Modelle zu untersuchen.

5. Erkenntnisse von Dritten.

Unsere Ergebnisse wurden während des Förderzeitraums mit Partnern aus den vom BMEL geförderten KI-Zucht und X-KIT präsentiert und diskutiert (siehe Abschnitt II.6). Jedoch sind während der Projektlaufzeit von dritter Seite keine für die Durchführung des Vorhabens relevanten Ergebnisse bekannt geworden. Die intensivsten Diskussionen unserer Ergebnisse und der Gedankenaustausch fanden mit unseren Projektpartnern statt. Eine gemeinsame Publikation der Ergebnisse wird derzeit zwischen den Projektpartnern diskutiert (siehe Abschnitt II.6). Auf der Grundlage dieser wissenschaftlichen Treffen und einer ständigen Suche in der aktuellen Literatur können wir feststellen, dass unsere Methoden und Ergebnisse des maschinellen Lernens nach unserem besten Wissen aktuell zu den innovativsten gehören und während des Förderzeitraums nicht von anderen Mitgliedern des KI-Zucht-Konsortiums oder den BMEL-geförderten KI und X-KIT Projektes untersucht bzw. übertroffen wurden. Einige wenige Teile des Projekts werden zur Zeit von anderen nationalen und internationalen Gruppen erforscht, jedoch überwiegend für andere Kulturarten als Raps.

6. Veröffentlichungen.

Obwohl wir die Forschungsarbeiten von Meilenstein M18 und M19 in der Hälfte der ursprünglich geplanten Zeit abgeschlossen haben, bietet der hohe Neuheits- und

Innovationsgrad der Ergebnisse aus dem KI-Zucht Arbeitspaket 4 eine hervorragende Möglichkeit, unsere Ergebnisse - gemeinsam mit unseren KI-Zucht Partnern - mittelfristig in hochrangigen wissenschaftlichen Zeitschriften zu veröffentlichen. Im Rahmen des KI-Forschungsvorhabens wurden unsere Forschungsergebnisse auch auf den Symposien der vom BMEL geförderten KI-Zucht und X-KIT Projekte vorgestellt, die ebenfalls eine gute Plattform für den wissenschaftlichen Austausch, das fachliche Feedback und mögliche zukünftige Kooperationen mit anderen KI-Forschungsvorhaben boten. Dementsprechend führen wir im Folgenden die Informationen auf, die der Präsentation und Veröffentlichung der Ergebnisse des KI-Zucht Arbeitspaket 4 entsprechen.

Publikationen:

- In Vorbereitung: Lozada-Chávez I., Krenzer D., Beltran V. F., Malenica M., Abbadi A., Snowdon R.J. and Bogdan M. (2025). **SNP-evolutionarily informed deep-learning enhances the power of predictive rapeseed phenotypes from environmental data.**

Präsentationen:

Abschlussveranstaltung der BMEL-geförderten KI-Projekte und X-KIT am Fraunhofer IESE, 25-26 Apr 2024, Kaiserslautern, Germany

- L. Roscher-Ehrig, I. Gabur, I. Lozada-Chávez, F. Beltran Velandia, D. Krenzer, M. Bogdan, M. Malenica, A. Abbadi, R. Snowdon: **Project KI-Zucht: Atrificial intelligence as the basis for future plant breeding based on “Big Data”.**