



Projektarbeit zur Teilnahme am Wettbewerb „Jugend forscht“ 2023/24

Ester, Säure oder Amin?

— Auswertung von IR-Spektren durch künstliche Intelligenz

Fachbetreuer: Herr Alexander Rübsam

Projektbetreuer: Frau Dr. Marion Moor

Fachbetreuer Informatik: Herr Detlef Flock

Name: Albert-Valentin Meyer (10spa)

Paula Kaltwasser (10spa)

Joshua Schraud (10spa)

Abgabe: 17. März 2024, Erfurt

Projektübersicht

Dieses Projekt befasst sich hauptsächlich mit der informatischen Auswertung typischer IR-Spektren organischer Verbindungen. Hierfür wurden eigene neuronale Netze konstruiert und trainiert. Dies erfolgte in Python.

Zunächst wurde ein genaueres Kenntnis sowohl über künstliche neuronale Netze an sich, als auch über die physikalischen Vorgänge innerhalb der IR-Spektroskopie und somit das Zustandekommen von IR-Spektren gewonnen.

Hiernach erfolgten sowohl die Konstruktion der neuronalen Netze, als auch das Sammeln der Daten zum Training der neuronalen Netze. Die Trainingsdaten wurden händisch aus IR-Spektren öffentlich zugänglicher Datenbanken wie der „SDBS“ (*Spectral Database for Organic Compounds* des japanischen *National Institute of Advanced Industrial Science and Technology*) zusammengestellt. Hierbei musste allerdings eine deutliche Reduktion der Komplexität der Daten erfolgen, sodass ein optimales Training der Netze bei gleichbleibender Aussagekraft möglich wurde.

Die Konstruktion der neuronalen Netze erfolgte in der Programmiersprache Python und resultierte in zwei unterschiedlichen Ansätzen. Einer der Ansätze wurde unter Benutzung des Moduls „Tensorflow“ realisiert, der andere ohne selbiges. Infolgedessen zeigten beide Ansätze, bei gleichlangem Training, mehrmals variierende Genauigkeiten. Ebenso wurden die verschiedenen Einflüsse variierender Lernraten und Lerndauer auf die Genauigkeit der Netze untersucht und graphisch dargestellt.

Letztendlich konnten mit beiden Modellen relativ hohe Genauigkeiten erreicht werden; das in Tensorflow realisierten neuronalen Netz erlangte hier 89%, während der andere Ansatz ebenso beachtliche 84% erreichen konnte. Dies gelang bei einer Größe des Datensatzes von etwa 50 IR-Spektren organischer Verbindungen. Der Programmcode lässt sich unter <https://github.com/JoshuaSchraud/Jugend-forscht-2024-Meyer-Schraud-Kaltwasser> [letzter Zugriff: 11.01.2024] abrufen.

Insgesamt lässt sich also feststellen, dass mit den von uns konstruierten neuronalen Netzen bereits einige Vorhersagen mit relativ hoher Genauigkeit durchführen lassen. Um die statistische Relevanz dieser Ergebnisse zu verbessern, könnte zukünftig vor allem die Größe der zum Training verwendeten Datensätze erhöht werden. Dies wäre beispielsweise über eine Automatisierung des Sammelns und Konvertierens der Daten möglich.

Inhaltsverzeichnis

| | | |
|----------|--|-----------|
| 1 | Fachliche Kurzfassung | 4 |
| 2 | Motivation und Fragestellung | 5 |
| 3 | Hintergrund und theoretische Grundlagen | 6 |
| 3.1 | Funktionsweise künstlicher neuronaler Netze | 6 |
| 3.2 | Aufbau und Arbeitsweise eines künstlichen neuronalen Netzes | 6 |
| 3.3 | Training des Netzes mittels eines Datensatzes | 7 |
| 3.4 | Vorteile künstlicher neuronaler Netze gegenüber herkömmlichen Algorithmen | 8 |
| 4 | Funktionsweise der IR-Spektroskopie | 9 |
| 4.1 | Grundlagen der Spektren elektromagnetischer Strahlung | 9 |
| 4.2 | Darstellung der energetischen Zustände im Molekül | 9 |
| 4.3 | Auswertung eines IR-Spektrums | 10 |
| 5 | Vorgehen und Methoden zur Vorbereitung und zum Training des neuronalen Netzes | 11 |
| 5.1 | Zielstellung sowie Zusammenstellung der Datensätze | 11 |
| 5.2 | Konvertierung der Daten | 12 |
| 5.3 | Konstruktion und Training des neuronalen Netzes | 14 |
| 6 | Ergebnisse und Ergebnisdiskussion der Trainings- und Testdurchläufe | 15 |
| 6.1 | Vergleich der Ergebnisse verschiedener Lösungsansätze | 15 |
| 6.2 | Kritische Reflexion der Ergebnisse und des angewandten Verfahrens | 17 |
| 6.3 | Zukünftige Nutzungs- und Verbesserungsmöglichkeiten | 17 |
| 7 | Fazit und Ausblick | 18 |

1 Fachliche Kurzfassung

Der Aufbau künstlicher neuronaler Netze orientiert sich an der Struktur biologischer Gehirne. Mittels miteinander verknüpfter Neuronen können Signale gerichtet weitergegeben werden. Diese Neuronen werden bei künstlichen neuronalen Netzen in Schichten angeordnet. Damit ein neuronales Netz für seine vorgesehene Funktion einsetzbar wird, muss es vorher einem Training unterzogen werden. Eine mögliche Methode hierfür ist die „Backpropagierung“.

Das physikalische Analyseverfahren der IR-Spektroskopie basiert auf den unterschiedlich starken Wechselwirkungen von Infrarotstrahlung mit diversen organischen sowie anorganischen Molekülen. Die von den Molekülen absorbierte Strahlung wird in diskreten Bewegungszuständen „gespeichert“. Diese Absorptionsvorgänge lassen sich in verschiedenen quantenmechanischen Modellen darstellen. Durch diese Modelle lassen sich die Absorptionsbanden der IR-Spektren bereits gut nachvollziehen.

Aus den spezifischen Absorptions- bzw. Emissionsspektren können die Strukturmerkmale unbekannter Substanzen umkehrbar eindeutig hergeleitet werden. Hierfür scheint eine künstliche Intelligenz ideal. Weil die Verarbeitung graphischer Daten durch künstliche neuronale Netze jedoch häufig mit einem großen Zeitaufwand verbunden ist, da eine enorme Rechenleistung nötig ist, erschien das Training eines neuronalen Netzes mittels graphisch vollständigen IR-Spektren als ungeeignet.

Infolgedessen musste die Komplexität der durch die IR-Spektren vorgegebenen Daten passend zum Aufbau des neuronalen Netzes reduziert werden. Deshalb wurden die Graphen vereinfacht, in dem zwischen allen relevanten Absorptionsbanden der IR-Spektren lineare Funktionen gezeichnet wurden. Die einzelnen Punkte dieses vereinfachten Graphen wurden dann an das neuronale Netz übergeben.

Wir entschieden uns für zwei verschiedene Lösungsansätze bei der Konstruktion des neuronalen Netzes. In einem der beiden verwendeten wir das Framework „Tensorflow“, im anderen erfolgte dies nicht. Der erstere der beiden Ansätze erwies sich mit einer Genauigkeit der Vorhersage der Strukturmerkmale von ca. 89 % als effektiver. Dafür wurde das neuronale Netz 60 Epochen lang trainiert, da darüber hinausgehendes Training das Problem des „Overfitting“ mit sich bringt. Mit dem zweiten Ansatz konnte lediglich eine Genauigkeit von circa 84 % erreicht werden. Somit ist es gelungen, ein neuronales Netz zu erschaffen, dass anhand IR-Spektren Strukturmerkmale von Molekülen mit recht großer Treffsicherheit ableitet.

2 Motivation und Fragestellung

Das perfekte Analyseverfahren ist der Traum eines jeden Chemikers. In den letzten 20 Jahren leistete die Infrarotspektroskopie, kurz IR-Spektroskopie, einen wichtigen Beitrag zum Erreichen dieses Ziels.

Mithilfe physikalischer Analyseverfahren, wie der IR-Spektroskopie, kommt man dem komplexen Ziel, einen Analyt ohne genauere Kenntnisse seiner Zusammensetzung zu identifizieren, erheblich näher. Zwar macht dieser Methode heutzutage die sogenannte NMR-Spektroskopie Konkurrenz, allerdings ist die IR-Spektroskopie bis heute in ihrer Effektivität und Kosteneffizienz ungeschlagen.

Aufgrund ihrer nützlichen Eigenschaften wird jene Art der quantitativen Analyse heutzutage weiterhin für zahlreiche Anwendungen eingesetzt. So dient die IR-Spektroskopie beispielsweise der Strukturaufklärung unbekannter bzw. nicht-isomerenreiner Verbindungen oder dem Vergleich von Testsubstanzen mit bekannten Verbindungen.

Nach dem Einsatz eines IR-Spektroskops erhält man ein stoffklassen- bzw. substanzspezifisches IR-Spektrum. Beispielsweise ist in der organischen Chemie eindeutig bestimmbar, ob es sich um einen Alkohol oder eine völlig andere funktionelle Gruppe handelt.

Da die Auswertung von Spektren ein gewisses Maß an Fachkunde und Zeitaufwand benötigt und dadurch durchaus fehlerbehaftet ist, entschieden wir uns, diesen Vorgang zu automatisieren.

Um dies erfolgreich umzusetzen, nutzten wir ein künstliches neuronales Netz. Diese Art der künstlichen Intelligenz (KI) spielt gerade in den letzten zehn Jahren in der Öffentlichkeit eine vermehrt relevante Rolle. Sie simulieren die Arbeitsweise eines biologischen Gehirns und sind daher sehr lernfähig und in der Lage, komplexe Probleme, an denen herkömmliche Algorithmen scheitern, zu lösen.

Daher war unser Ziel, die Chemie und die Informatik zu vereinen und ein Programm zur Auswertung von IR-Spektren mithilfe eines neuronalen Netzes zu entwickeln. Letzteres wollten wir so trainieren, dass es in der Lage sein würde, die Strukturmerkmale von organischen Substanzen anhand ihres IR-Spektrums zu identifizieren.

Für seine tatkräftige Unterstützung im Fachbereich Chemie danken wir Herrn Alexander Rübsam. Weiterhin möchten wir uns bei Herrn Detlef Flock bedanken, welcher uns nützliche Hinweise im Fachbereich Informatik gab und bei Fragen stets zur Seite stand. Ein besonderer Dank gilt zudem Herrn Elias Nolik Hölzer, welcher uns, falls wir einmal an unsere Grenzen im Bezug auf die KI stießen, ebenfalls unterstützte. Ebenso bedanken wir uns sehr herzlich bei Herrn Dr. Uwe Köhn, da er uns einige selbstaufgezeichnete IR-Spektren zur Verfügung stellte und uns die Möglichkeit gab, eigene IR-Spektren aufzuzeichnen.

3 Hintergrund und theoretische Grundlagen

3.1 Funktionsweise künstlicher neuronaler Netze

3.2 Aufbau und Arbeitsweise eines künstlichen neuronalen Netzes

Menschen fällt es leicht, große Menge an Informationen aus Bildern erfolgreich zu verarbeiten. Ohne große Bemühungen können wir den Inhalt erfassen und einordnen, was zu sehen ist. Eine solche Aufgabe fällt einem Computer unglaublich schwer. Stattdessen besitzt er die Fähigkeit, Rechenoperationen in einer für Menschen unerreichbaren Geschwindigkeit durchzuführen. Trotz ihrer hohen Komplexität und Leistungsfähigkeit fehlt Computern die Intelligenz, um für einen Menschen einfach scheinenden Aufgaben zu bewältigen. Um dieses Problem zu lösen, konstruierte man künstliche neuronale Netze, man modellierte sie nach der Struktur biologischer Neuronen. Wie in einem biologischem Gehirn kommt es anschließend zu komplexen Verknüpfungen dieser Neuronen. Solche Neuronen übertragen ein elektrisches Signal, welches von Neuron zu Neuron immer weitergegeben wird.¹

Wie in Abbildung 1 zu sehen ist, bestehen künstliche neuronale Netze auch aus einzelnen Neuronen, welche als Knoten bezeichnet werden. Sie ordnen sich in Schichten, den Layern, an. Die Eingabeschicht ist die erste Schicht an Neuronen, die die gegebenen Eingangssignale entgegennimmt. Anschließend folgt eine beliebige Anzahl an versteckten Schichten. Die Ausgabeschicht gibt dann die ausgerechneten Werte an den Benutzer des Programms zurück. Die einzelnen Neuronen sind dabei mit jedem Knoten der vorherigen und der nachfolgenden Schicht verbunden, sodass sich ein künstliches neuronales Netz ergibt.²

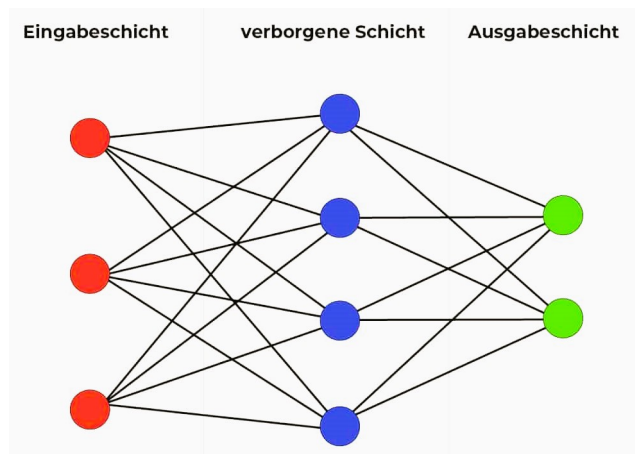


Abbildung 1: Schema eines künstlichen neuronalen Netzes

Die Verbindungen zwischen den Neuronen haben eine bestimmte Gewichtung. Je stärker dieses Gewicht einer Verbindung ist, desto größer ist der Einfluss eines Knotens auf einen anderen. Die Gewichte werden während des Lernprozesses verändert, wodurch das neuronale Netz seine Ausgaben verbessert. Wenn nun das Netz so komplex ist, dass es Knoten im Übermaß für die Lösung eines Problems besitzt, werden die zusätzlichen Verbindungen durch das Verringern des Gewichts abgeschwächt. Diese Verbindungen besitzen fast keinen Einfluss mehr auf das Ergebnis des Netztes, da ihre Leistung nicht benötigt wird.³

¹RASHID, TARIQ: *Neuronale Netze selbst programmieren*; O'Reilly; 1. Auflage; Paderborn; 2017, S. 2, S. 30

²Lüdemann, Jonah; *Grundlagen Neuronaler Netze [PDF]*; S. 3, S. 5; https://wr.informatik.uni-hamburg.de/_media/teaching/sommersemester_2020/siw-20_luedemann_grundlagen-nn_ausarbeitung.pdf letzter Zugriff: 23.04.2023

³Cociancig, Christina; *Künstliche neuronale Netze - Aufbau & Funktionsweise*; <https://www.jaai.de/post/kuenstliche-neuronale-netze-aufbau-funktion> letzter Zugriff: 23.04.2023

Diese Wirkung der Gewichte wird deutlich, wenn die Funktionsweise jedes einzelnen Knotens genauer betrachtet wird. Die Neuronen künstlicher neuronaler Netze funktionieren analog zu biologischen Neuronen. Ein biologisches Neuron gliedert sich in drei Abschnitte: Zellkörper, Dendriten und Axon. Die Dendriten können Signale anderer Nervenzellen aufnehmen, die über das Axon weitergeleitet werden, welches in vielen Verzweigungen endet. Diese Verzweigungen knüpfen daraufhin wieder an andere Neuronen an. Jedoch wird erst dann ein Signal in nachfolgende Neuronen weitergeleitet, wenn das Signal stark genug ist. Dementsprechend funktioniert auch ein künstliches Neuron. Beliebige viele Eingänge werden zunächst mit den entsprechenden Gewichten der Verbindungen multipliziert und anschließend alle addiert. Nachfolgend wird auf die Summe eine Aktivierungsfunktion angewandt. Diese bewirkt, dass das Neuron nur feuert, wenn das Signal groß genug ist; anderenfalls wird das Signal unterdrückt.

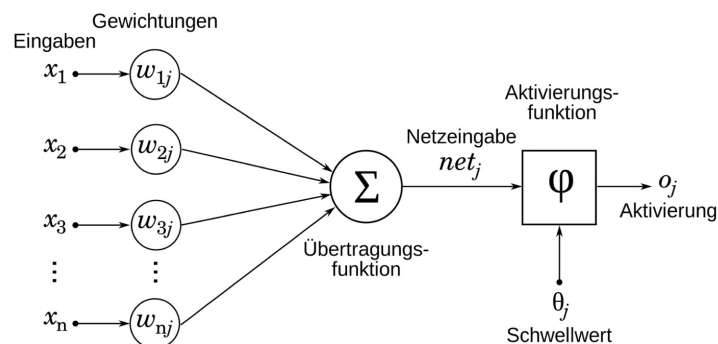


Abbildung 2: Schema zur Funktionsweise eines Neurons künstlicher neuronaler Netze

3.3 Training des Netzes mittels eines Datensatzes

Damit ein künstliches neuronales Netz funktionieren kann, muss es mit einem geeigneten Datensatz trainiert werden. Hierbei wird die Methode der Backpropagation angewandt; diese wird ebenfalls als Fehlerrückführung bezeichnet. Das Verfahren besteht aus zwei Schritten. Im ersten Schritt wird der Fehler des künstlichen neuronalen Netzes berechnet, anschließend werden im zweiten Schritt die Verbindungsgewichte so angepasst, dass der Fehler möglichst minimiert wird. Man nutzt den Einfluss der Gewichte auf die Ausgaben jedes Neurons. Ein Datensatz besteht dabei aus vielen verschiedenen Eingabedaten und der erwarteten Ausgabe. Die Eingabedaten werden dem Netz übergeben, sodass es eine Ausgabe für diese errechnet. Die erhaltene Ausgabe wird nun mit dem Zielwert verglichen und die Differenz zwischen dem Zielwert und dem erhaltenen Wert wird gebildet. Diese Differenz wird als Fehler bezeichnet. Der Fehler des gesamten Netzes wird anschließend auf jeden einzelnen Knoten jeder Schicht aufgeteilt.⁴

Wenn das neuronale Netz den Fehler aller Neuronen berechnet hat, müssen die Gewichte so angepasst werden, dass das neuronale Netz Ergebnisse mit geringeren Fehlern produziert. Hierfür wird das sogenannte Gradientenabstiegsverfahren verwendet. Dazu wird der Fehler in Abhängigkeit der Gewichte des

⁴Bayerischer Rundfunk; *Nervenzelle - Aufbau und Funktion*; <https://www.br.de/telekolleg/faecher/biologie/biologie-2-nervenzelle100.html> letzter Zugriff: 23.04.2023

neuronalen Netzes betrachtet. Um herauszufinden, welches Gewicht für jede einzelne Verknüpfung der Neuronen nötig ist, damit der Fehler möglichst klein wird, soll der Anstieg dieser Funktion möglichst minimiert werden. Wenn man also beispielsweise einen positiven Anstieg erhält, muss er sich verringern; wenn der Anstieg der Fehlerfunktion negativ ist, soll er vergrößert werden. Hierfür muss man den Anstieg der Fehlerfunktion von dem alten Verknüpfungsgewicht subtrahieren, um das neue Gewicht zu erhalten. Davor wird der Anstieg aber noch mit der Lernrate α multipliziert, die die Stärke der Veränderung des Gewichtes moderiert. Der Anstieg der Fehlerfunktion ist das partielle Differenzial des Fehlers E dividiert durch das partielle Differenzial des Gewichts. Der Anstieg kann durch Differenzierung berechnet werden. So erhält man folgenden Term zur Berechnung des neuen Gewichts:⁵

$$w_{neu} = w_{alt} - \alpha \cdot \frac{\partial E}{\partial w_{alt}} \quad (1)$$

Dadurch wird das Gewicht beispielsweise bei einem großen, positiven Anstieg verhältnismäßig stark verändert. Der zu minimierende Anstieg wird mit zunehmender Anzahl an Trainingsbeispielen immer kleiner und so die Veränderungen des Gewichtes immer feiner. Mit jedem einzelmem Trainingsdatenelement werden die Gewichte so angepasst, dass der Fehler des neuronalen Netzes immer kleiner wird und das Netz immer genauere Vorhersagen treffen kann.⁶

3.4 Vorteile künstlicher neuronaler Netze gegenüber herkömmlichen Algorithmen

Da die künstlichen Neuronen neuronaler Netze in ihrem Aufbau biologischen Neuronen ähneln, sind sie in der Lage, intelligente Lernansätze zu verfolgen und somit eine neuartige Herangehensweise an verschiedene Probleme zu erlernen. Herkömmliche Algorithmen bestehen aus einer Abfolge elementarer Anweisungen, die Schritt für Schritt abgearbeitet werden. Sie sind daher im Gegensatz zu neuronalen Netzen weniger flexibel, können auf teilweise unvorhersehbare, noch nicht berücksichtigte Einflüsse schlecht reagieren und erzielen somit keine sinnvollen Ergebnisse.

Neuronale Netze hingegen können mittels geeigneter Trainingsdatensätze immer genauere Ergebnisse berechnen, komplexere Zusammenhänge erkennen und Prognosen erstellen. Sie sind so in der Lage, sich selbst zu verbessern, was bei „normalen“ Algorithmen nach jedem Durchlauf durch einen Programmierer oder interne Routinen erfolgen muss. Dies stellt einen Vorteil für die verschiedensten Anwendungsbereiche dar. Jedoch kann das Sammeln von Daten sehr aufwendig sein und wenn dieser Datensatz fehlerhaft ist, kann es dazu führen, dass das neuronale Netz nach unerwünschten Kriterien sortiert oder den Datensatz nur auswendig lernt.⁷

⁵RASHID, TARIQ: *Neuronale Netze selbst programmieren*; O'Reilly; 1. Auflage; Paderborn; 2017, S.73, S. 81, S. 86

⁶siehe 8

⁷Jonah Lüdemann; *Grundlagen Neuronaler Netze*; https://wr.informatik.uni-hamburg.de/_media/teaching/sommersemester_2020/siw-20_luedemann_grundlagen-nn_ausarbeitung.pdf, S. 8, letzter Zugriff: 08.05.2023

4 Funktionsweise der IR-Spektroskopie

4.1 Grundlagen der Spektren elektromagnetischer Strahlung

Wie der Name „Infrarotspektroskopie“ verrät, beruht dieses physikalische Analyseverfahren auf dem Einsatz von Licht im infraroten Bereich. Es handelt sich also um elektromagnetische Strahlung.

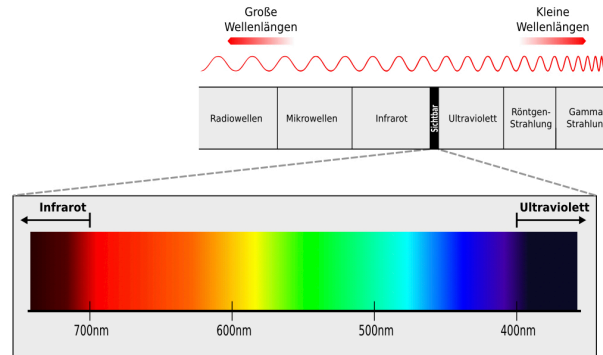


Abbildung 3: Bereiche elektromagnetischer Strahlung

Mit einer Wellenlänge von 2.500 nm - 15.400 nm liegt die Infrarotstrahlung über dem Bereich des sichtbaren Lichts, jedoch im Bereich der Schwingungsniveaus einiger Molekülbindungen bzw. Atome oder Atomgruppen. Die Moleküle absorbieren die Strahlung, was zu einer Schwingungsanregung der Bindungen führt. Die dafür notwendigen Energien bzw. Frequenzen hängen vom jeweiligen Molekül ab und sind je nach vorhandener funktioneller Gruppe verschieden. In der IR-Spektroskopie wird die unterschiedliche Absorbanz (A) dieser Molekülgruppen gemessen. Im Graphen wird diese später als Transmittanz (Transmittanz = 10^{-A-2} , T in %) ⁸ dargestellt. Insgesamt wird also aufgezeichnet, welche Gruppen bei diskreten Wellenzahlen die IR-Strahlung stärker absorbieren.

4.2 Darstellung der energetischen Zustände im Molekül

Wird ein Molekül mit Licht einer diskreten Wellenlänge λ bestrahlt, kommt es je nach Molekül bei einigen Wellenlängen zur Aufnahme von Energie. Dabei kann zwischen dem Ausgangszustand E_1 und dem angeregten, energiereicheren Zustand E_2 unterschieden werden. ⁹ In der Spektroskopie wird der energiereichere Zustand (') üblicherweise mit einem und der energieärmere Zustand (") mit zwei Strichen bezeichnet.

Der Zusammenhang zwischen den beiden Zuständen, der Energiedifferenz und der Frequenz ν des absorbierten Lichts, ist in Gleichung (4) ¹⁰ wiedergegeben:

$$\Delta E = E' - E'' = h \cdot \nu \quad (2)$$

⁸Sigma-Aldrich: *Transmittance to Absorbance Table*; <https://www.sigmaaldrich.com/DE/de/technical-documents/technical-article/analytical-chemistry/photometry-and-reflectometry/transmittance-to-absorbance>, letzter Zugriff: 10.05.2023

⁹TU-Clausthal: *IR SPEKTROSKOPIE*; <https://www.pc.tu-clausthal.de/fileadmin/homes/praktikum/05-IR-SPEKTROSKOPIE.pdf>, S. 2, letzter Zugriff: 08.05.2023

¹⁰siehe Fußnote 12

Die Energie im angeregtem Zustand E'' kann auf unterschiedliche Weisen gespeichert werden. So treten im Molekül beispielsweise Rotationsbewegungen oder Schwingungen bzw. Vibrationen auf. Die Energie wird also zum Teil als Rotationsenergie ($E_{Rot.}$) oder Schwingungsenergie ($E_{Vib.}$) aufgenommen. Ebenfalls könnte die Ladungsverteilung um die Atomkerne verändert, also die Elektronen angeregt werden, wobei die Energie in Form von $E_{El.}$ gespeichert wird. Allerdings muss dies hier aus zwei Gründen nicht berücksichtigt werden:

Einerseits ist die Elektronenanregungsenergie wesentlich höher als $E_{Rot.}$ und $E_{Vib.}$. Andererseits kommt es im Wellenzahlbereich der IR-Spektroskopie (400 cm^{-1} bis 4000 cm^{-1}) überhaupt nicht zum Erreichen dieser elektrisch angeregten Zustände, da die eingestrahlte Energie zu gering ist.

Da $E_{El.}$ vernachlässigt werden kann, lässt sich die Gesamtenergie des Zustandes E'' näherungsweise folgendermaßen darstellen ¹¹:

$$E' = E_{Rot.} + E_{Vib.} \quad (3)$$

Üblicherweise wird in der Spektroskopie anstatt der Wellenlänge λ , die Wellenzahl $\bar{\nu}$ (mit $\bar{\nu} = \frac{1}{\lambda}$) verwendet, da ein direkt proportionaler Zusammenhang zur eingestrahlten Energie besteht (da $\nu = \frac{c}{\lambda}$, also $\nu = c \cdot \bar{\nu}$).

4.3 Auswertung eines IR-Spektrums

Aus der Messung mit einem IR-Spektrometer ergibt sich ein stoffspezifisches IR-Spektrum. An der y-Achse eines solchen Spektrums ist der Transmissionsgrad in Prozent aufgeführt. Dieser gibt an, wie viel des Infrarotlichts die Stoffprobe durchdringt. Liegt die Transmission bei 100 Prozent, passt die Frequenz der Strahlung zu keiner Schwingungsfrequenz der Atome oder Bindungen im Molekül. Den Bereich dieser entsprechenden Wellenzahl (bzw. Wellenlänge) zeigt die x-Achse. ¹²

Je nachdem, wie gering der Transmissionsgrad und dementsprechend hoch die Absorption ist, ergeben sich im Spektrum unterschiedliche Tiefpunkte, die im Folgenden als Absorptionsbanden bezeichnet werden. Diese sind je nach funktioneller Gruppe unterschiedlich tief bzw. breit.

Eine weitere Besonderheit stellt die sogenannte „Fingerprint-Region“ dar. Dieser Begriff bezeichnet im IR-Spektrum den Bereich mit Wellenzahlen unter 1500 cm^{-1} . ¹³ Zwar lassen sich in dieser Region keine eindeutigen Signale funktioneller Gruppen feststellen, da es hier zu Interferenzen der zahlreich auftretenden Deformations- und Streckschwingungen kommt. Jedoch sind die dort auftretenden Muster substanzspezifisch und können zur Identifizierung unbekannter Substanzen dienen. Für die Zwecke dieser Arbeit soll die Fingerprint-Region jedoch weitestgehend unbeachtet bleiben, da es das Ziel ist, Strukturmerkmale und keine konkreten Substanzen festzustellen. ¹⁴

¹¹TU-Clausthal; *IR SPEKTROSKOPIE*; <https://www.pc.tu-clausthal.de/fileadmin/homes/praktikum/05-IR-SPEKTROSKOPIE.pdf>, S. 2, letzter Zugriff: 08.05.2023

¹²Bruker Corporation: *FT-IR Spektroskopie Grundlagen*; <https://www.bruker.com/de/products-and-solutions/infrared-and-raman/ft-ir-routine-spectrometer/what-is-ft-ir-spectroscopy.html>, letzter Zugriff: 10.05.2023

¹³Scilearn Sidney: *Infrared: Fingerprint-Region*, <https://scilearn.sydney.edu.au/OrganicSpectroscopy/?type=Infrared&page=Fingerprint%20Region>, letzter Zugriff: 10.05.2023

¹⁴Direct Industry: *FT-IR-Spektrometer INVENIO Series*; <https://www.directindustry.de/prod/bruker-optics-gmbh-co-kg/product-57234-2055385.html>, letzter Zugriff: 17.04.2023

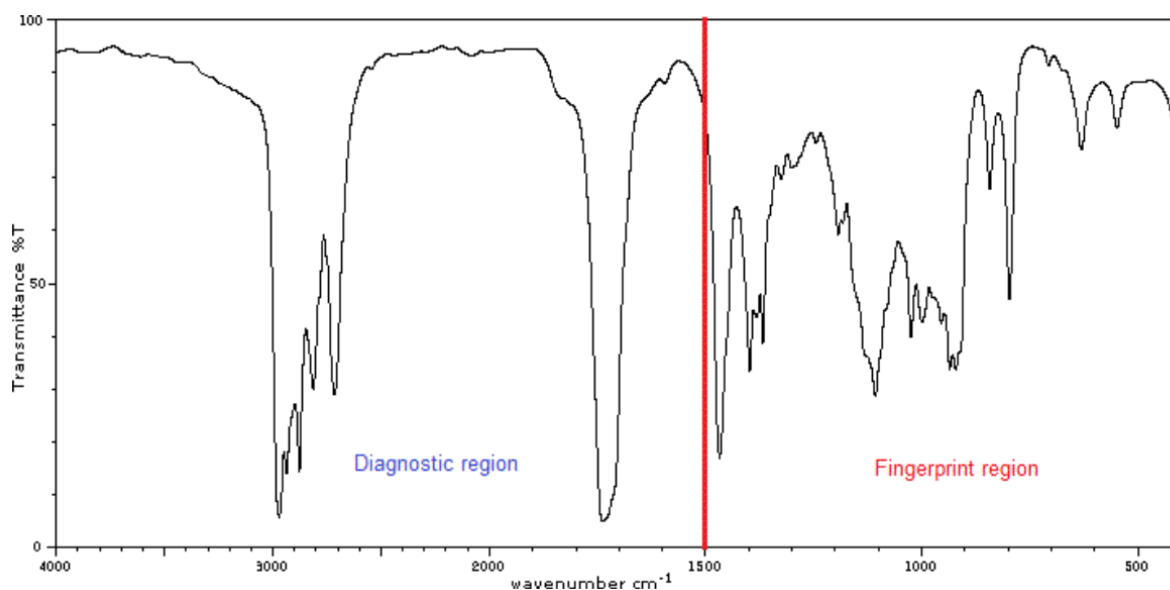


Abbildung 4: beschriftetes Beispiel IR-Spektrum

5 Vorgehen und Methoden zur Vorbereitung und zum Training des neuronalen Netzes

5.1 Zielstellung sowie Zusammenstellung der Datensätze

Zunächst musste das generelle Vorgehen geplant werden. So handelte es sich beim ursprünglichen Ziel darum, die Daten durch das neuronale Netz aus einem Bild, der typischen Form der IR-Spektren, herauszulesen und die Spektren anschließend den korrekten Strukturmerkmalen zuordnen zu lassen. Dies stellte sich jedoch als sehr zeitaufwändig heraus und könnte zudem simpler durchgeführt werden. Daher war die neue Überlegung, nicht das gesamte Spektrum, sondern bloß die Absorptionsbanden zu verwenden, um das Netz zu trainieren. Die Fingerprint-Region wurde zunächst mit einbezogen.

Anschließend musste die Zuordnung der Strukturmerkmale erfolgen und jedes der dreizehn Strukturmerkmale erhielt eine eigene Kategorie:

| <i>Kategorie (Stellung)</i> | <i>Strukturmerkmal</i> | <i>Darstellung</i> |
|-----------------------------|------------------------|---------------------------|
| 1 | Alkane | C-H |
| 2 | Alkine | C≡C sowie C-H |
| 3 | Aromaten | C-H (Ringschwingung etc.) |
| 4 | Stickstoff | N-H |
| 5 | Amid | N-H |
| 6 | Cyanid | C≡N |
| 7 | Alkene | C=C |
| 8 | Cycloalkene | C=C |
| 9 | Alkohole | O-H |
| 10 | Aldehyde | H-C=O |
| 11 | Ketone | C=O |
| 12 | Ester | COOC |
| 13 | Carbonsäuren | COOH |

Tabelle 1: Kategorien und Strukturmerkmale ¹⁵

Die IR-Spektren bezogen wir von der „Spectral Database for Organic Compounds, SDBS“ ¹⁶ des „National Institute of Advanced Industrial Science and Technology, AIST“ aus Japan. Die SDBS ist eine öffentliche Datenbank, welche nach eigenen Angaben Molekülspektren von über 54.100 organischen Verbindungen enthält. Für die jeweiligen Spektren waren unter den Visualisierungen zwar die Absorptionsbanden angegeben, mussten jedoch händisch in CSV-Dateien (*Comma-Separated-Values-Dateien*) übertragen werden. Bei CSV-Dateien handelt es sich um eine spezielle Form der Textdateien. Diese sind plattformunabhängig und eignen sich, um große Datensätze in einem einfach zugänglichem Format zu speichern ¹⁷. Die Daten werden über Kommata getrennt und liegen in einer Art Tabelle vor. Die Kategorien der Verbindungen (Strukturmerkmale) werden hier durch „0“ oder „1“ dargestellt. Ist das Strukturmerkmal vorhanden, steht in der Datei an der entsprechenden Stelle (Kategorie 1-13) eine Eins, falls nicht, eine Null. Daraus ergibt sich an den ersten 13 Stellen in jeder Zeile der „Tabelle“ (CSV-Datei) eine Folge von Nullen und Einsen, wobei auch mehrere Einsen verschiedene Kategorien kennzeichnen können. Eine Carbonsäure weist also an der 13. Stelle (siehe Tabelle 1) und auch an der 9. bzw. 10. Stelle eine Eins auf, da sie ebenfalls die Strukturmerkmale eines Alkohols bzw. Aldehyds besitzt.

5.2 Konvertierung der Daten

Als wir die Daten der „SDBS“ in eine CSV-Datei übertrugen, mussten sie so konvertiert werden, dass sie gut als Eingangsdaten für das künstliche neuronale Netz fungieren konnten. Daher entschieden wir uns, die Absorptionsbanden jedes Spektrums mit linearen Funktionen zu verbinden und dem Netz die Koordinaten jedes Punktes mit ganzzahligem x-Wert zu übermitteln, sodass pro Molekül 3600 Punkte entstanden. Dies hat seine Ursache darin, dass die meisten IR-Spektroskopien im Bereich der Wellenzahl von 400 bis 4000 cm^{-1} gemessen werden und die Daten der „SDBS“ auch in diesem Bereich liegen.¹⁸

¹⁵Universität Mainz: OC2 Blatt 12, Elektromagnetische Wellen und Spektroskopie <https://download.uni-mainz.de/ak-detert/oc2blatt12.pdf>, S. 4, zuletzt besucht: 09.08.2023

¹⁶AIST: https://sdb.sdb.aist.go.jp/sdb/cgi-bin/cre_index.cgi, zuletzt besucht: 05.09.2023

¹⁷Marr, Tobias: Was ist eine .csv-Datei?; <https://www.heise.de/tipps-tricks/Was-ist-eine-csv-Datei-6321660.html>, zuletzt besucht: 10.05.2023

¹⁸AK Kappenberg; Infrarotspektroskopie Grundlagen; <https://www.kappenberg.com/experiments/spektren/pdf-aka11/y20.pdf>, S. 1, letzter Zugriff: 09.05.2023

Wir wählten dieses Vorgehen, da so eine einheitliche Größe von 7200 benötigten Eingabeneuronen gegeben und der Graph des tatsächlichen IR-Spektrums dem Graphen der Eingabedaten so ähnlich war. Diese Ähnlichkeit ist am Beispiel des Spektrums von Ethanol erkennbar (siehe Abbildung 5). Bei Spektrum (1) handelt es sich um das Ursprungsspektrum aus der SDBS; Spektrum (2) wurde nach erläuterten Verfahren konvertiert.

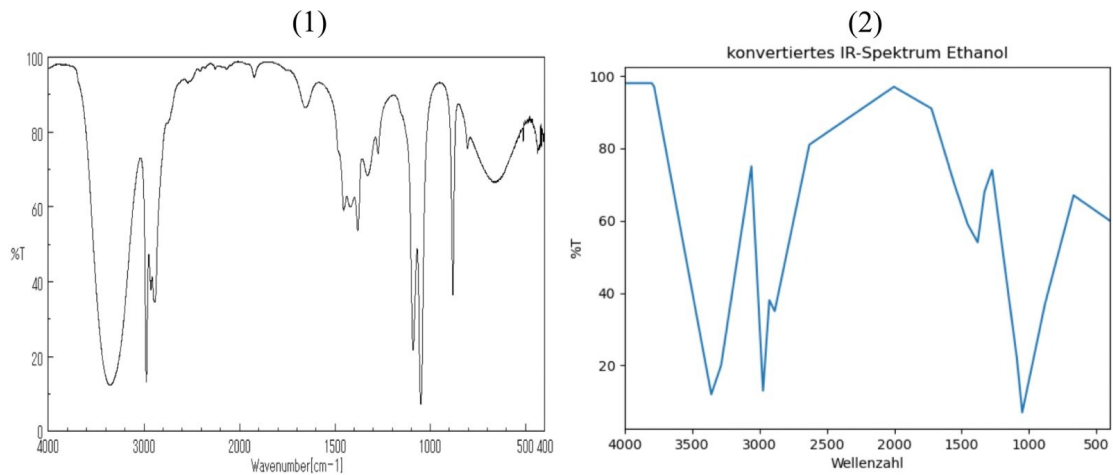


Abbildung 5: Gegenüberstellung der Spektren von Ethanol

Da im Datensatz des „AIST“ nur die relevanten Absorptionsbanden abgebildet waren, sehen die Graphen dieser vereinfachten und bereinigten Darstellung etwas anders aus (siehe Abbildung 6).

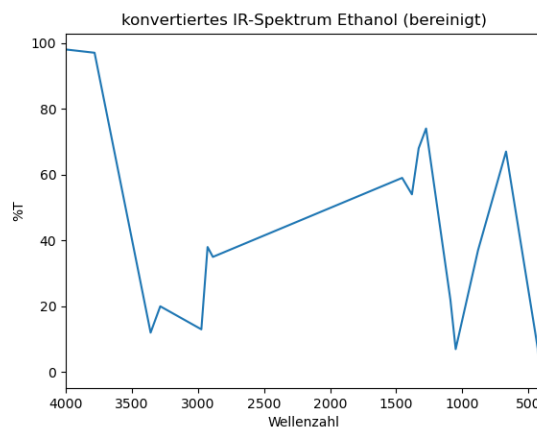


Abbildung 6: konvertiertes, bereinigtes Spektrum von Ethanol

Bereinigt bedeutet hier, dass nur die relevanten Absorptionsbanden angegeben und die regulären Punkte mit einer sehr starken Transmittanz entfernt wurden. Diese sind für die Berechnungen irrelevant.

5.3 Konstruktion und Training des neuronalen Netzes

Die Konstruktion des Netzes erfolgte in Python, unter Verwendung von Tensorflow, einem Framework für maschinelles Lernen, sowie Keras, einer Deep-Learning-Bibliothek, ebenfalls in Python geschrieben. Für jede Kategorie wurde ein eigenes, sogenanntes „Sequential-Modell“ erstellt, welches erlaubt, das Netz eigenständig Schicht für Schicht aufzubauen.¹⁹ Somit besteht das gesamte Konstrukt aus 13 einzelnen neuronalen Netzen, die zusammen dazu fähig sind, den IR-Spektren ihre Strukturmerkmale zuzuordnen. Jedes neuronale Netz baut sich folgendermaßen auf: Nach einer Eingabeschicht mit 7200 Neuronen folgte die sogenannte „Flatten-Layer“. Diese sorgt dafür, dass mehrdimensionale, in diesem Fall zweidimensionale, Eingaben in eindimensionale Daten umgewandelt werden. Anschließend folgen zwei „Hidden-Layers“ mit jeweils 32 und 64 Neuronen, auch „Nodes“ oder Knoten. Wie auch bei der Eingabeschicht handelt es sich hierbei um „Dense-Layer“. Eine Dense-Layer ist eine Schicht, in welcher alle Neuronen mit jedem Neuron der vorherigen Schicht verbunden sind²⁰. An diese drei versteckten Schichten schließt sich eine letzte Schicht, die Ausgabeschicht, an. Diese besitzt 2 Knoten, wobei immer einer den Wert 0 und einer den Wert 1 annimmt. Je nachdem, welcher der beiden Knoten den Wert 1 annimmt, ordnet das neuronale Netz das Strukturmerkmal dem Spektrum zu oder nicht. An jedes neuronale Netz werden die Testdaten einzeln übergeben, sodass sich aus den Antworten jedes Netzes eine Gesamtanalyse des IR-Spektrums ergibt.

Das hier beschriebene Verfahren wurde gewählt, da es eine bessere Nachvollziehbarkeit bzw. eine einfachere Konstruktion ermöglicht. Zwar mag es sich hierbei nicht um die effektivste Lösung handeln, dennoch ist feststellbar, dass dieses Verfahren die gewünschten Ergebnisse erzielt. Des Weiteren wurde ein weiterer mittels Tensorflow programmierter Ansatz erprobt. Nach diesem Modell wurde ein Netz erstellt, welches alle 13 Kategorien mit jeweils 13 Knoten in der Ausgabeschicht bestimmen sollte. Da sich dabei aber Fehler mit Tensorflows Errechnung der Genauigkeit ergaben, wurde dieser Ansatz nicht weiter verfolgt. Neben der Nutzung des Moduls Tensorflow wurde auch ein Programm entworfen, das die gefragte Leistung ohne Tensorflow erzielen soll.

Für das detaillierte Training der neuronalen Netze wurde das in 2.2 beschriebene Trainingsverfahren angewandt. Dazu wurden zunächst ein Trainings- und ein Testdatensatz erstellt. Im Trainingsdatensatz befinden sich die Daten und zugehörige Kategorien von Verbindungen, anhand derer das Netz lernen sollte, diese zuzuordnen. Wenn Zuordnungen nicht korrekt sind, werden die Verbindungsgewichte zwischen den Neuronen angepasst. Im Testdatensatz hingegen befinden sich dem Netz unbekannte Daten. Diese werden nach den jeweiligen Trainingsdurchläufen (*Epochen*) durchgegangen. Anhand des Testdatensatzes soll nun festgestellt werden, ob das Netz die korrekten Zuordnungen durchführt. Die Anzahl der Epochen und die Lernrate können angepasst werden, um die Genauigkeit des Netzes zu optimieren. Die Auswirkung der Änderung der Epochenanzahl soll ebenso, wie die jeweils erreichte Genauigkeit des Netzes festgestellt werden. Im Trainingsdatensatz befinden sich 51 Verbindungen, im Testdatensatz hingegen neun.

¹⁹Chollet, François; *Das sequentielle Modell*; https://keras.io/guides/sequential_model/, letzter Zugriff: 10.05.2023

²⁰Verma, Yugesh; *A Complete Understanding of Dense Layers in Neural Networks*; <https://analyticsindiamag.com/author/yugesh-vermaanalyticsindiamag-com/>, letzter Zugriff: 12.05.2023

6 Ergebnisse und Ergebnisdiskussion der Trainings- und Testdurchläufe

6.1 Vergleich der Ergebnisse verschiedener Lösungsansätze

Von den drei angewandten Verfahren erwies sich das Verfahren, welches eine Konstruktion aus 13 separat arbeitenden neuronalen Netzen darstellte, die jeweils binäre Ausgaben lieferten, als besonders effektiv. Dies äußerte sich im Erreichen der bisher höchsten Genauigkeit sowie einer effektiven Fehlerminimierung:

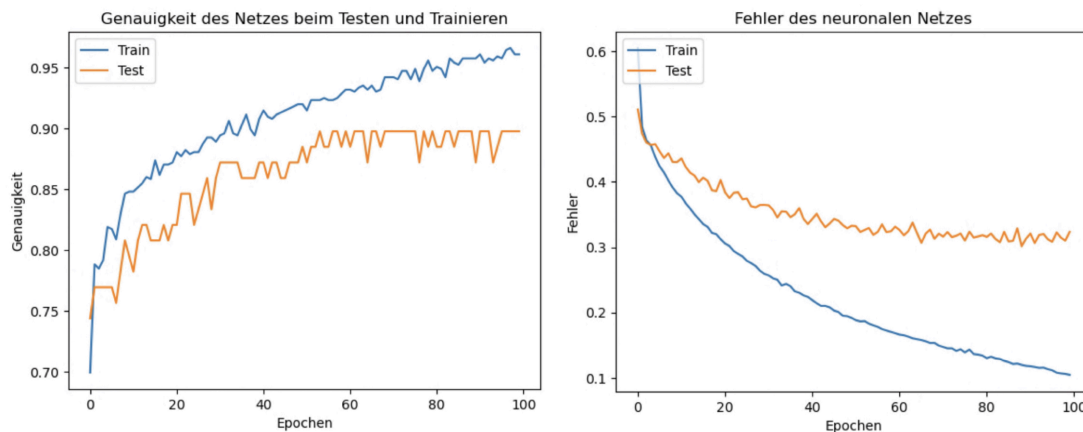


Abbildung 7: Genauigkeit und Fehler im Trainings- sowie im Testdurchlauf der neuronalen Netze in Abhängigkeit der Epochen

Wie im linken Graphen zu sehen ist, bleibt die Testgenauigkeit ab circa 60 Epochen, abgesehen von einigen Schwankungen, konstant. Währenddessen steigt die Genauigkeit für die Trainingsdaten weiter an und der Abstand zwischen Trainings- und Testgenauigkeit wird größer. Dies ist auf ein Phänomen zurückzuführen, welches sich „Overfitting“ oder auch Überanpassung, nennt. Beim Overfitting lernt das Netz prinzipiell die Trainingsdaten „auswendig“ und spezialisiert sich auf das grundlegende Muster, lernt bzw. erfasst es jedoch nicht.²¹ Dadurch steigt die Genauigkeit beim Testen der Trainingsdaten, während sich das Netz aber tatsächlich nicht mehr verbessert. Allerdings gilt zu beachten, dass der sogenannte „best fit“ noch nicht stark überschritten wurde, da sonst zu beobachten wäre, dass die Geraden noch stärker divergieren würden. Ähnliches ist auch bei der Fehlerfunktion zu beobachten (Abbildung 7, rechter Graph): Ab ungefähr 60 Epochen wird der Fehler für die Trainingsdaten minimiert, während der für die Testdaten annähernd gleich bleibt. Aufgrund dessen entschieden wir uns, das Netz nur mit 60 Epochen zu trainieren.

wZusätzlich zur Anzahl der Epochen kann auch die Lernrate verändert werden, um die größtmögliche Leistung zu erzielen. Für das aus 13 separaten Netzen bestehende neuronale Netz ergab sich folgender Graph:

²¹Alexander Thamm GmbH; *Overfitting*; <https://www.alexanderthamm.com/de/data-science-glossar/overfitting/>, letzter Zugriff: 24.05.2023

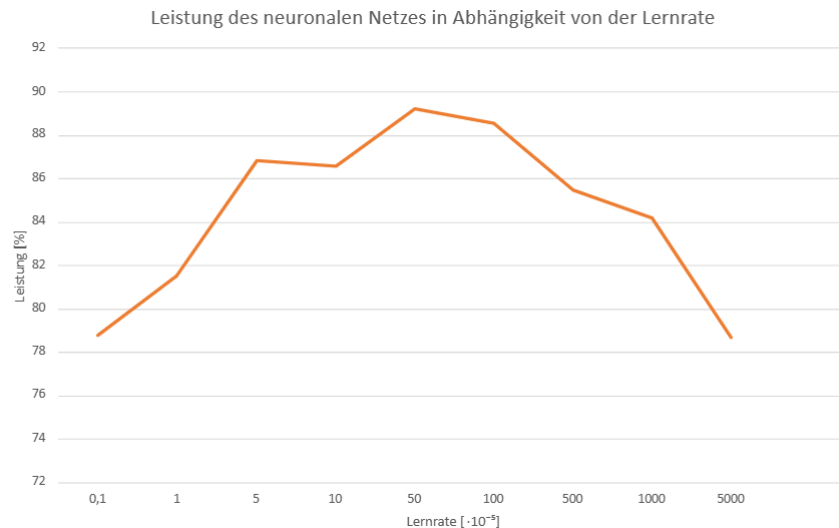


Abbildung 8: Genauigkeit im Testdurchlauf der neuronalen Netze in Abhängigkeit von der Lernrate

Hier zeigt sich, dass bei einer Lernrate von 0,0005 die höchsten Genauigkeitswerte resultieren. So war es mithilfe dieses Ansatzes möglich, mit 60 Epochen und einer Lernrate von 0,0005 eine konsistente Genauigkeit von circa 89 % zu erzielen. Diese schwankte bei erneuter Ausführung des Programms um einige Prozentpunkte.

Die Vorhersagen des neuronalen Netzes ohne den Einsatz des Moduls Tensorflow waren etwas weniger zutreffend als die des ersten Ansatzes, sodass die erreichte Genauigkeit um einige Prozentpunkte geringer war:

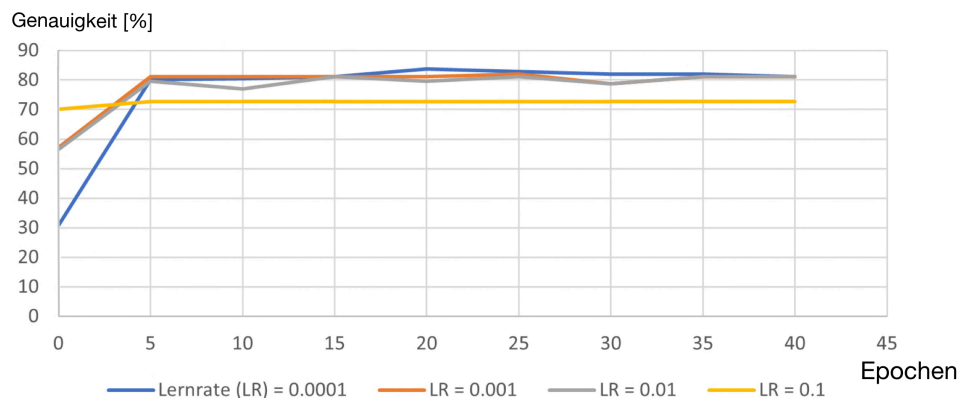


Abbildung 9: Genauigkeit im Testdurchlauf des neuronalen Netzes ohne Tensorflow in Abhängigkeit der Epochen bei verschiedenen Lernraten

Bei den in Abbildung 9 gezeigten Genauigkeiten ist vor allem auffällig, dass die Leistungen bei unterschiedlichen Lernraten und Epochen nur noch geringfügig abweichen. Eine mögliche Ursache hierfür könnte die deutlich geringere Komplexität des Netzes im Vergleich zum anderen Ansatz mittels Tensorflow sein. Dieses neuronale Netz besteht nur aus drei Schichten, während das mit Tensorflow erstellte Netz ganze fünf Schichten enthält. Dadurch hat es geringere Fähigkeiten zur Erkennung von Zusammenhängen und Mustern in den Daten, sodass die Genauigkeit bei weiterem Training konstant bleibt. Auch die Zeit, die das Programm benötigt, um einmal durchzulaufen, ist wichtig für die Bewertung des

Ergebnisses. Diesbezüglich schneidet das ohne Tensorflow erstellte neuronale Netz deutlich besser ab, als das mit Tensorflow erstellte. Während das Netz, welches ohne Tensorflow erstellt wurde, nur ca. 11 Sekunden brauchte, benötigte das mit Tensorflow erstellte Netz für einen Trainings- und Testdurchlauf ca. 100 Sekunden. Damit braucht das leistungsfähigere Netz mehr als neunmal so lang wie das andere neuronale Netz. Die Gründe hierfür liegen wahrscheinlich in der höheren Komplexität des Netzes, welches mittels Tensorflow erstellt wurde, und darin, dass beim leistungsfähigeren Ansatz 13 einzelne Netze trainiert und getestet wurden. Die höchsten Leistungen des Ansatzes ohne Tensorflow konnten mit einer Lernrate von 0.0001 bei 20 Epochen erreicht werden. Für unbekannte Moleküle betrug die Genauigkeit der Vorhersagen ca. 83,76 %. Wie auch beim ersten Ansatz kann es Abweichungen von einigen Prozentpunkten geben.

Zusammenfassend lässt sich aussagen, dass beide dieser Ansätze eine solide und brauchbare Leistung erbringen konnten. Bei der beschränkten Anzahl an Datenwerten, die den Netzen zur Verfügung standen, konnte das vor allem durch den Einsatz mehrerer Schichten respektive durch eine hohe Komplexität des Netzes erreicht werden. Dies könnte einer der Gründe sein, weshalb das neuronale Netz, welches mittels Tensorflow umgesetzt wurde, deutlich bessere Werte erzielen konnte. Trotz der größeren Trainings- und Testdauer handelt es sich unserer Meinung nach bei dem mit Tensorflow erstellten neuronalen Netz um das nützlichere und vielversprechendere, da für uns die größere Genauigkeit ein höheres Gewicht für die Bewertung der Leistungsfähigkeit eines neuronalen Netzes darstellt.

6.2 Kritische Reflexion der Ergebnisse und des angewandten Verfahrens

Im Zuge der Projektarbeit gelang es uns, sowohl ein neuronales Netz als auch ein komplexeres Konstrukt aus 13 neuronalen Netzen zu erschaffen, welche jeweils IR-Spektroskopen auswerten und so die Strukturmerkmale des Moleküls ausgeben. Dafür haben wir eigene Datensätze auf Grundlage der Daten der „SDBS“ erstellt. Insgesamt umfassen diese die Absorptionsbanden von über 50 Molekülen aus verschiedenen Stoffgruppen der organischen Chemie. Unter den Molekülen für die Datensätze waren Ester, Ether, Alkohole, Carbonsäuren, Amide, Nitrile, Aromaten, Alkene, Alkine und Aldehyde. Darüber hinaus waren wir erfolgreich dabei, diese Daten so zu konvertieren, dass sie als Eingabe für das neuronale Netz dienen können.

Allerdings ging mit dem Erstellen des Datensatzes ein großer Zeitaufwand einher, weswegen sein Umfang eher gering war. Dennoch sind die erhaltenen Daten aussagekräftig und die Leistung des neuronalen Netzes ist durchaus hoch. Damit das Netz aber tatsächlich im Berufsalltag einsetzbar ist, braucht es eine größere Anzahl an Trainingsdaten.

6.3 Zukünftige Nutzungs- und Verbesserungsmöglichkeiten

Trotz der bereits erläuterten Hindernisse, gibt es dennoch Möglichkeiten, die Zuverlässigkeit und damit die Anwendbarkeit des neuronalen Netzes zu verbessern. So bestünde die Möglichkeit bzw. die Notwendigkeit, den Trainings- sowie Testdatensatz zu erweitern. Um den bereits erwähnten Zeitaufwand zu minimieren, wäre eine Automatisierung auf Grundlage von Bilderkennung lohnenswert. Überdies wäre es möglich, die Effizienz des Programms zu verbessern, indem beispielsweise nicht 13 separate neuronale Netze, sondern nur ein einziges verwendet würde. Um eine Einbindung des hier entwickelten neuronalen Netzes in einen konkreten Service zu erreichen, bedürfte es weiterer Entwicklung. Dennoch wäre ein breites Anwendungsspektrum denkbar.

7 Fazit und Ausblick

Künstliche Intelligenzen wie neuronale Netze spielen in der heutigen Zeit vermehrt eine wichtige Rolle, sowohl im alltäglichen Leben, als auch in der Wissenschaft. Sie unterstützen den Menschen in verschiedenen Bereichen, sei es bei der Bilderkennung von Smartphones oder bei Früherkennungssystemen, wie sie in einigen modernen Fahrzeugen integriert sind.²²

Aufgrund dieses breiten Anwendungsbereichs, planten wir, ein neuronales Netz für unsere Projektarbeit zu nutzen. Wir setzten uns zum Ziel, ein neuronales Netz mittels geeigneter Trainingsdatensätze so zu konzipieren, dass es in der Lage sein würde, Diagramme von IR-Spektren verlässlich auszuwerten.

Die IR-Spektroskopie ist ein physikalisches Analyseverfahren, welches mittels Infrarotstrahlung zur Strukturaufklärung unbekannter Substanzen in der chemischen Analytik verwendet wird. Dieses Verfahren bietet die Chance, Substanzen zu identifizieren und Aussagen über im Molekül vorhandene funktionelle Gruppen zu machen. Um diese Aussagen jedoch treffen zu können, müssen die Diagramme, welche sich bei einer solchen Messung ergeben, ausgewertet werden. Dies geschieht hauptsächlich durch den Menschen und ist daher oft fehlerbehaftet.

Um den Alltag im Labor zu erleichtern und möglichst verlässliche Aussagen über die molekulare Struktur unbekannter Stoffe treffen zu können, sollte die Auswertung der IR-Spektren von einem eigens konstruierten, künstlichen neuronalen Netz übernommen werden.

Für diesen Zweck konstruierten wir ein neuronales Netz in „Python“, welches anschließend mithilfe der Methode der Backpropagation trainiert wurde. Dafür erstellten wir einen Trainingsdatensatz, dessen Inhalte wir von einer öffentlichen Datenbank für Molekülspektren bezogen. Ebenso wurde ein Testdatensatz aus den Spektrendaten organischer Verbindungen zusammengestellt. Die Strukturmerkmale, welche vom Netz errechnet werden sollten, wurden in 13 Kategorien eingeteilt und dem Datensatz beigelegt. Diese Daten konvertierten wir in eine CSV-Datei und übergaben sie dem Netz.

Jedoch musste das Konvertieren der Daten händisch vorgenommen werden, weshalb die Erstellung der Datensätze viel Zeit beanspruchte. Infolgedessen war der Umfang des Datensatzes relativ gering. Dennoch gelang es, eine verhältnismäßig hohe Leistung durch das neuronale Netz zu erzielen. Mit einem Konstrukt aus 13 separaten Netzen, welche jeweils auf die Erkennung eines Strukturmerkmals ausgelegt waren, wurde eine Höchstleistung von rund 89% erreicht.

Durch eine abschließende kritische Reflexion der erlangten Ergebnisse ist erkennbar, dass immer noch ein gewisses Verbesserungspotenzial besteht. So könnte das aufwändige Konvertieren der Daten mittels einer Bilderkennung beschleunigt und so der Umfang der Datensätze zum Trainieren und Testen erhöht werden. Ebenfalls würde die Konstruktion eines einzigen Netzes, welches die Tätigkeit des von uns entwickelten Konstruktes an Netzen übernehmen könnte, eine gewisse Zeitersparnis mit sich bringen. Des Weiteren könnte die Effektivität und damit die Anwendbarkeit des Netzes, durch weitere Entwicklung so erhöht werden, dass es tatsächlich im Labor einsetzbar wäre.

²²Luber, Stefan; Litzel, Nico; Was ist ein Neuronales Netz?; <https://www.bigdata-insider.de/was-ist-ein-neuronales-netz-a-686185/>, letzter Zugriff: 10.05.2023

Quellen- und Literaturverzeichnis

Gedruckte Literaturquellen

- [1] RASHID, TARIQ: *Neuronale Netze selbst programmieren*; O'Reilly; 1. Auflage; Paderborn; 2017
- [2] GÜNZLER, H.; GREMLICH, H.-U: *IR-Spektroskopie (4. Auflage)*; Wiley-VCH; 4. Auflage; Weinheim; 2003

Internetquellen

- [3] AK Kappenberg: *Infrarotspektroskopie Grundlagen* <https://www.kappenberg.com/experiments/spektren/pdf-aka11/y20.pdf>, zuletzt abgerufen: 09.05.2023
- [4] Aunkofer, Benjamin: *Fehler-Rückführung mit der Backpropagation* <https://data-science-blog.com/blog/2019/01/29/fehler-ruckfuhrung-mit-der-backpropagation/>, zuletzt abgerufen: 08.05.2023
- [5] Bayerischer Rundfunk: *Nervenzelle - Aufbau und Funktion* <https://www.br.de/telekolleg/faecher/biologie/biologie-2-nervenzelle100.html>, zuletzt abgerufen: 08.05.2023
- [6] Bruker Corporation: *FT-IR Spektroskopie Grundlagen* <https://www.bruker.com/de/products-and-solutions/infrared-and-raman/ft-ir-routine-spectrometer/what-is-ft-ir-spectroscopy.html>, zuletzt abgerufen: 17.04.2023
- [7] Chollet, François: *Das sequentielle Modell* https://keras.io/guides/sequential_model/, zuletzt abgerufen: 10.05.2023
- [8] Cociancig, Christina: *Künstliche neuronale Netze - Aufbau und Funktionsweise* <https://www.jaai.de/post/kuenstliche-neuronale-netze-aufbau-funktion>, zuletzt abgerufen: 23.04.2023
- [9] DEEPLIZARD: *Keras-Tensorflow Tutorial*; <https://deeplizard.com/learn/video/RznKVRTFkBY>, zuletzt abgerufen: 08.05.2023
- [10] Direct Industry: *FT-IR-Spektrometer INVENIO Series* <https://www.directindustry.de/prod/bruker-optics-gmbh-co-kg/product-57234-2055385.html>, zuletzt abgerufen: 10.05.2023
- [11] Direct Industry: *FT-IR-Spektrometer INVENIO Series* <https://www.directindustry.de/prod/bruker-optics-gmbh-co-kg/product-57234-2055385.html>, zuletzt abgerufen: 10.05.2023
- [12] Hezel, Nico : *KNN: Vorteile und Nachteile*; <https://data-science-blog.com/blog/2015/09/12/knn-vorteile-und-nacheile/>, zuletzt abgerufen: 01.02.2023

- [13] Institut für Physikalische Chemie, TU Clausthal: *IR SPEKTROSKOPIE*; <https://www.pc.tu-clausthal.de/fileadmin/homes/praktikum/05-IR-SPEKTROSKOPIE.pdf>, zuletzt abgerufen: 17.04.2023
- [14] Lehmann, Thomas: *IR - Spektroskopie* https://www.bcp.fu-berlin.de/chemie/chemie/studium/ocpraktikum/_Unterlagen_Spektroskopie/ir.pdf, zuletzt abgerufen: 08.05.2023
- [15] Lindner, Jean-Pierre; Nolte, Thies: *FT-IR Spektroskopie*; https://www.uni-muenster.de/imperia/md/content/physikalische_chemie/praktikum/ft_ir_spektroskopie_lindner_jean_pierre__thies__nolte__thies_.pdf, zuletzt abgerufen: 17.04.2023
- [16] Luber, Stefan; Litzel, Nico: *Was ist ein Neuronales Netz?*; <https://www.bigdata-insider.de/was-ist-ein-neuronales-netz-a-686185/>, zuletzt abgerufen: 02.02.2023
- [17] Lüdemann, Jonah: *Grundlagen Neuronaler Netze* https://wr.informatik.uni-hamburg.de/_media/teaching/sommersemester_2020/siw-20_luedemann_grundlagen-nn_ausarbeitung.pdf, zuletzt abgerufen: 23.04.2023
- [18] Marr, Thobias: *Was ist eine .csv-Datei?* <https://www.heise.de/tipps-tricks/Was-ist-eine-csv-Datei-6321660.html>, zuletzt abgerufen: 10.05.2023
- [19] Mwit, Derrick: *Keras-Metrics Programmierung*; <https://neptune.ai/blog/keras-metrics>, zuletzt abgerufen: 10.05.2023
- [20] National Institute of Advanced Industrial Science and Technology: https://sdb.sdb.aist.go.jp/sdb/cgi-bin/cre_index.cgi, zuletzt abgerufen: 05.09.2023
- [21] Sigma-Aldrich: *Transmittance to Absorbance Table* <https://www.sigmaaldrich.com/DE/de/technical-documents/technical-article/analytical-chemistry/photometry-and-reflectometry/transmittance-to-absorbance>, zuletzt abgerufen: 10.05.2023
- [22] Scilearn Sidney: *Infrared: Fingerprint-Region* <https://scilearn.sydney.edu.au/OrganicSpectroscopy/?type=Infrared&page=Fingerprint%20Region>, zuletzt abgerufen: 10.05.2023
- [23] Universität Mainz: *OC2 Blatt 12, Elektromagnetische Wellen und Spektroskopie* <https://download.uni-mainz.de/ak-detert/oc2blatt12.pdf>, zuletzt abgerufen: 09.08.2023
- [24] Universität Ulm: *Grundlagen und Aufbau von neuronalen Netzen* <https://www.informatik.uni-ulm.de/ni/Lehre/SS02/Evosem/OptimierungKNNFolien.pdf>, zuletzt abgerufen: 08.05.2023
- [25] Verma, Yugesh; *A Complete Understanding of Dense Layers in Neural Networks* <https://analyticsindiamag.com/author/yugesh-vermaanalyticsindiamag-com/>, zuletzt abgerufen: 12.05.2023
- [26] Zheng, Liqing; Wu, Na: *INFRAROT-SPEKTROSKOPIE* https://analytik.ethz.ch/praktika/phys_anal/IR/IR_Anleitung.pdf, zuletzt abgerufen: 08.05.2023

Abbildungen und Tabellen

- [27] Abbildung 1: Schema eines künstlichen neuronalen Netzes; <https://datasolut.com/neuronale-netzwerke-einfuehrung/>, zuletzt abgerufen: 16.04.2023
- [28] Abbildung 2: Graphische Darstellung der Sigmoid-Funktion; <https://de.wikipedia.org/wiki/Sigmoidfunktion#/media/Datei:Sigmoid-function-2.svg>, zuletzt abgerufen: 23.04.2023
- [29] Abbildung 3: Bereiche elektromagnetischer Strahlung; <https://www.leifiphysik.de/optik/elektromagnetisches-spektrum/grundwissen/sichtbares-licht/>, zuletzt abgerufen: 16.04.2023
- [30] Abbildung 4: beschriftetes Beispiel IR-Spektrum; <https://chemistryscore.com/how-to-interpret-ir-spectra/>, zuletzt abgerufen: 18.04.2023
- [31] Abbildung 5: Gegenüberstellung der Spektren von Ethanol; eigene Darstellung
- [32] Abbildung 6: konvertiertes, bereinigtes Spektrum von Ethanol; eigene Darstellung
- [33] Abbildung 7: Genauigkeit und Fehler im Trainings- sowie im Testdurchlauf der neuronalen Netze in Abhängigkeit der Epochen; eigene Darstellung
- [34] Abbildung 8: Genauigkeit im Testdurchlauf der neuronalen Netze in Abhängigkeit von der Lernrate; eigene Darstellung
- [35] Abbildung 9: Genauigkeit im Testdurchlauf des neuronalen Netzes ohne Tensorflow in Abhängigkeit der Epochen bei verschiedenen Lernraten; eigene Darstellung

Softwarequellen

- [36] Jupyter: JupyterLab: A Next-Generation Notebook Interface; <https://jupyter.org/>, zuletzt abgerufen: 18.04.2023
- [37] NumPy: Installing NumPy; <https://numpy.org/install/>, zuletzt abgerufen: 18.04.2023
- [38] Matplotlib.pyplot: Installation; <https://matplotlib.org/stable/users/installing/index.html>, zuletzt abgerufen: 18.04.2023
- [39] SciPy: Installation, <https://scipy.org/install/>, zuletzt abgerufen: 18.04.2023
- [40] TensorFlow: Install TensorFlow 2; <https://www.tensorflow.org/install>, zuletzt abgerufen: 18.04.2023

Unterstützungsleistungen

Inhaltlich wurden wir von Herrn Alexander Rübsam, Lehrer für Chemie am Spezialschuleteil des ASG Erfurt, durch Hilfe bei inhaltlichen Fragen bezüglich der IR-Spektroskopie ebenso wie quantenmechanischen Modellen unterstützt. Diese Unterstützung fand über die gesamte Projektdauer statt.

Des Weiteren fand eine Bereitstellung von Material durch Herrn Detlef Flock, Lehrer für Informatik am Spezialschuleteil des ASG Erfurt, über die gesamte Projektdauer statt. Er stellte uns die Quelle [1] als institutionsinterne Literaturquelle bereit und unterstützte uns mit der Beantwortung von Fragen bezüglich LaTeX bzw. formalen Vorgaben.

Außerdem erhielten wir einige Unterstützung durch Herrn Dr. Uwe Köhn, welcher uns selbst ermittelte IR-Spektren zur Verfügung stellte und uns des weiteren ermöglichte IR-Spektren von selbst synthetisierten Substanzen aufzuzeichnen und auszuwerten.

Ebenso unterstützte uns Herr Elias N. Hölzer, Schüler am Spezialschuleteil des ASG Erfurt, mehrfach durch technische Hilfe. So fand Unterstützung bei der Findung von Ansätzen zur Beseitigung von Fehlerquellen im Programmcode für den Lösungsansatz mittels Keras statt und einige fachliche Fragen konnten beantwortet werden. Für all diese Unterstützung bedanken wir uns herzlich.

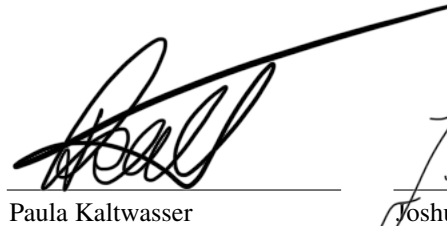
Eidesstattliche Erklärung

Hiermit versichern wir an Eides Statt, dass wir die vorliegende Arbeit selbstständig verfasst, keine anderen als die angegebenen Quellen und Hilfsmittel genutzt haben und alle Ausführungen, die anderen Schriften wörtlich oder sinngemäß entnommen wurden, kenntlich gemacht haben.

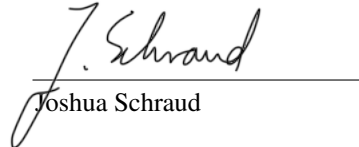
Erfurt, 17. März 2024



Albert-Valentin Meyer



Paula Kaltwasser



Joshua Schraud