Albert-Ludwigs-Universität Freiburg

**Automatisierte Identifizierung von Wirkstoffen in wissenschaftlicher Literatur mithilfe maschineller Lernverfahren**

Bachelorarbeit

*Pharmazeutische Wissenschaften*

Manuel Dorer

Fakultät für Chemie und Pharmazie,

Institut für Pharmazeutische Wissenschaften,

Pharmazeutische Bioinformatik

vorgelegt von

Manuel Dorer

am XX .XX 2021, Freiburg

Erstgutachter: Prof. Dr. Stefan Günther

Pharmazeutische Bioinformatik,

Albert-Ludwigs-Universität Freiburg

Zweitgutachter: Jun.-Prof. Dr. Jennifer Andexer,

Pharmazeutische und Medizinische Chemie,

Albert-Ludwigs-Universität Freiburg

Betreuer: Ammar Qaseem,

Pharmazeutische Bioinformatik,

Albert-Ludwigs-Universität Freiburg



Eigenständigkeitserklärung

Name des Studenten: Manuel Dorer

Matrikelnummer: 4531053

*ERKLÄRUNG*

*zur Abgabe der Bachelor-Arbeit   
im Studiengang B.Sc. Pharmazeutische Wissenschaften*

Hiermit versichere ich, dass ich die Bachelorarbeit selbstständig verfasst und keine andere als die von mir angegebenen Quellen und Hilfsmittel benutzt habe. Alle Zitate sind gekennzeichnet und alle Abbildungen enthalten nur die originalen Daten und sind in keinem Fall inhaltsverändernder Bildbearbeitung unterzogen worden. Die abgegebene schriftliche und elektronische Fassung sind identisch. Weiterhin versichere ich, dass die Arbeit noch nicht anderweitig als Bachelorarbeit eingereicht wurde.

............................................ ........................................................

Ort/Datum Unterschrift des Studenten

# Abstract

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

# Inhaltsverzeichnis

[Abstract i](#_Toc65365131)

[Inhaltsverzeichnis ii](#_Toc65365132)

[1 Einleitung 1](#_Toc65365133)

[1.1 Begriffe aus dem Feld der KI 1](#_Toc65365134)

[1.1.1 KI 1](#_Toc65365135)

[1.1.2 maschinelles Lernen 2](#_Toc65365136)

[1.2 ~~KI in der Pharmazeutischen Bioinformatik~~ bessere überschrift vlt datamining 9](#_Toc65365137)

[1.3 Überblick über State of the Art von “ cpi ppi textmining“ 9](#_Toc65365138)

[1.4 Zielsetzung der Arbeit 9](#_Toc65365139)

[2 Motivation der Arbeit 10](#_Toc65365140)

[3 Material und Methoden 10](#_Toc65365141)

[3.1 Systemspecs 10](#_Toc65365142)

[3.2 Installation 10](#_Toc65365143)

[3.3 Datensätze 10](#_Toc65365144)

[3.4 Statistische Parameter zur Bewertung 10](#_Toc65365145)

[3.5 10-fache Kreuzvalidierung 10](#_Toc65365146)

[4 Ergebnisse 10](#_Toc65365147)

[5 Diskussion und Ausblick 11](#_Toc65365148)

[6 Zusammenfassung 11](#_Toc65365149)

[7 Literaturverzeichnis 12](#_Toc65365150)

# Einleitung

Am 15 März 2016 stand es fest: AlphaGo besiegt den zu der Zeit als stärksten Spieler der Welt geltenden Lee Sedol in einer Partie Go mit 4:1 X1X. Doch was hat dieses Ereignis mit autonom fahrenden Autos, der Wettervorhersage oder Spracherkennung zu tun?

Die Antwort auf diese Frage ist: In allen genannten Bereichen wird zunehmend mit künstlicher Intelligenz (KI) gearbeitet, bzw. gibt es immer mehr Ansätze, welche auf einer KI basieren X2-5X. KI’s kommen häufig dort zum Einsatz, wo große Datenmengen anfallen und zeitnah bewertet werden müssen. Dabei erkennen KI’s Muster in den Daten und können Vorhersagen über die weitere Entwicklung treffen, was für das Fällen von Entscheidungen von großer Bedeutung ist. Nicht zuletzt haben KI’s auch auf vielfältige Art und Weise ihren Weg in die naturwissenschaftliche Forschung gefunden. Anwendungsgebiete sind z.B.: Das Stellen von Hautkrebsdiagnosen in der Medizin X6X, Optimierung von Molekülsimulationen für die Werkstoffforschung in der Physik; ~~dem Finden von neuen potentiellen Wirkstoffen in der Pharmazie;~~ oder generell das Extrahieren von Wissen aus Daten. Besonders letzteres gewinnt im Zeitalter von Big Data immer mehr an Relevanz und soll in dieser Bachelorarbeit im Mittelpunkt stehen.

## Begriffe aus dem Feld der KI

In diesem Kapitel werden für die Arbeit wichtige Begriffe aus dem großen Themengebiet der KI erläutert. Dies stellt einerseits eine Basis für ein Grundverständnis dar und soll andererseits dazu dienen diese Arbeit, mit der verwendeten Methode, in das sehr weitläufige Gebiet der KI besser einzuordnen.

### KI

Eine einfache universelle Definition von KI gibt es nicht. Das liegt daran, dass der Begriff Intelligenz selbst nur schwer zu definieren ist. Die Fraunhofer-Gesellschaft beschreibt KI in einer Studie als „ein Teilgebiet der Informatik mit dem Ziel, Maschinen zu befähigen, Aufgaben »intelligent« auszuführen“ X7, s.8X Dabei wird versucht die menschliche Intelligenz bzw. die Entscheidungsstrukturen des Menschen mithilfe von Algorithmen und mathematischen Modellen in Computersystemen nachzubilden und diese auf eigene Problemstellungen anzuwenden. Im Zusammenhang mit KI wird zwischen schwacher und starker KI differenziert. Schwache KI’s können nur in einem Bereich ein konkretes Anwendungsproblem bewältigen. D.h. sie sind nicht in der Lage Erlerntes auf andere Bereiche zu übertragen. Typische Anwendungsgebiete sind: Schach/- GO-Computer, Bilderkennung, Spracherkennung, personalisierte Werbung und automatisierte Übersetzung.

Starke KI’s hingegen sind zurzeit nicht existent. Eine starke KI besitzt intellektuelle Fähigkeiten auf Augenhöhe eines Menschen. Diese beruhen darauf, dass die KI in der Lage ist erlerntes Wissen in andere Bereiche zu übertragen und auf neue Probleme anzuwenden. Ob und wann es eine solche allgemeine Intelligenz geben wird, ist in der Wissenschaft umstritten X8X. Der Streitpunkt ist dabei, inwiefern sich typisch menschliche Eigenschaften, wie das Vorhandensein eines Bewusstseins, mit einer KI vereinbaren lassen.

Die Anfänge in der Forschung und Entwicklung KI lassen sich auf das Jahr 1956 datieren, als am Dartmouth College eine Konferenz namens Dartmouth Conference stattfand. Einen ersten Boom verzeichneten KI’s in den 1980er Jahren über die sogenannten Expertensysteme. Die Entscheidungen, welche von den Expertensystemen getroffen werden, basieren auf einer langen Kette von Wenn-Dann-Verknüpfungen. Im zweiten KI-Winter sank das öffentliche Interesse an KI und damit auch die Gelder für die Forschung. Durch die Verwendung von maschinellem Lernen, einer weiteren, bedeutsameren Untergruppe der KI, erlebt diese in dem 21. Jahrhundert eine Blütezeit (siehe Abbildung 1).



Abbildung 1: Entwicklung der Publikationen im Bereich maschinelles Lernen für Deutschland (DE), Frankreich (FR), Großbritannien (UK), USA (US) und China (CN) sowie die restlichen EU-Staaten und übrige Länder  
Quelle: Frauenhofer…. S.

### maschinelles Lernen

Maschinelles Lernen stellt ein wichtiger Zweig von KI dar. Abbildung 2 veranschaulicht das grundlegende Prinzip des maschinellen Lernens.



Abbildung 2: Grundlegendes Prinzip des maschinellen Lernens; icon quelle

Ein IT-System, bestehend aus Lernalgorithmen und mathematischen Modellen, bekommt als Input einen Datensatz und ist in der Lage in diesem Muster und Zusammenhänge zu erkennen. Das erlernte Wissen wird von dem System genutzt, um selbständig ein Modell zu trainieren, welches eine mathematische Beschreibung der im Datensatz gefundenen Gesetzmäßigkeiten ist. Dieses Modell ist der Output des Systems und kann anschließend, auf für das System unbekannte Daten, angewendet werden, um neue Erkenntnisse zu gewinnen. Zusammenfassend kann man sagen, dass der große Unterschied von maschinellem Lernen zu herkömmlichem Programmieren ist, dass der Mensch dem IT‑System keine direkten Regeln vorgibt, nach denen ein bestimmter Output auszugeben ist, sondern das IT‑System diese selbständig finden muss.

Mit der Technik des maschinellen Lernens ist es möglich riesige Datenmengen effizient zu bearbeiten, was dem Menschen schlicht aufgrund der großen Menge nicht möglich wäre. Maschinelles Lernen lässt sich heutzutage in Bereichen wie der Bilderkennung oder der Textanalyse finden.

Ein viel zitiertes Beispiel ist der Einsatz in Antispamfiltern für E-Mails X9X. Hierfür wird ein Modell mit einem Trainingsdatensatz, bestehend aus vielen E-Mails und dem zugehörigen Label *Spam‑E‑Mail* oder *keine Spam‑E‑Mail*, trainiert. Das trainierte Modell ist dann in der Lage bei einer unbekannten E‑Mail zu erkennen, ob es sich um eine *Spam‑E‑Mail* handelt, oder nicht.

Die klassischen Schritte des maschinellen Lernens sind:

**Datenerfassung:** Überführen der Rohdaten, inklusive der Label, in eine maschinenlesbare Form.

**Datenaufbereitung:** Vorbereitung der Daten (z.B. Normalisieren und Durchmischen der Reihenfolge); Am Ende wird der fertige Datensatz auf einen getrennten Test- und Trainingsdatensatz aufgeteilt. Das Verhältnis in dem zufällig gesplittet wird ist standardmäßig 70 % Trainingsdatensatz zu 30 % Testdatensatz, kann aber auch variiert werden. Es ist wichtig sicherzustellen, dass die Verteilung der Daten zwischen den beiden erzeugten Datensätzen ähnlich ist. Eine Möglichkeit die Ähnlichkeit zu bewerten, ist der Vergleich der Anteile der identischen Label in beiden Datensätzen. Bei dem oben beschriebenen Antispamfilter, stellt man so sicher, dass der Anteil der Spam-E-Mails in beiden Datensätzen ähnlich groß ist. Es wird vermieden, dass das Modell einseitig trainiert wird und dadurch eine spätere Bewertung des Models weniger aussagekräftig werden würde.

**Auswahl des passenden Modells:** Je nach Anforderung, wird aus mehrenden Modellen ein passendes ausgewählt. In Abbildung 3**Fehler! Verweisquelle konnte nicht gefunden werden.** sind einige Modelle hierarchisch dargestellt. Grundsätzlich wird, wenn keine Labelung der Daten vorhanden ist, von unüberwachtem Lernen gesprochen. Der Antispamfilter ist ein Beispiel für überwachtes Lernen, da wie oben beschrieben, eine Labelung vorliegt. Ein überwachtes Lernen Modell wird mit gelabelten Daten trainiert, dass es zukünftig zu ungelabelten Daten ein Label als Output ausgeben kann. Innerhalb des überwachten Lernens wird zwischen den beiden Aufgabentypen Regression und Klassifizierung differenziert. Das Entscheidende einer Klassifizierung, zu der auch der Antispamfilter gehört, ist, dass die Anzahl möglicher Label eine endliche Zahl ist. Bei einer Regression hingegen weist das Modell dem Datenpunkt einen stetigen Wert zu, d.h. das Label hat unendlich viele Ausprägungen. Die für diese Arbeit wichtigen Modelle sind Kern‑Methodenund neuronale Netze. Bei beiden handelt es sich um überwachtes Lernen in Verbindung mit einer Klassifizierungsaufgabe.



Abbildung 3: Übersicht über verschiedene Modelle des maschinellen Lernens. Die für diese Arbeit relevanten Modelle sind in orangener Farbe dargestellt. Quelle(selber gemacht, vorlage aus dem Paper)

**Trainieren des Modells:** Beim Modelltraining geht wird das Modell mit dem Testdatensatz dahingehend optimiert, dass es später möglichst gut mit den unbekannten Daten zurechtkommt. Es kann beim Trainieren des Modells zu den in Abbildung 4 gezeigten Szenarien kommen.



Abbildung 4: Unteranpassung *(underfit),* gute Anpassung *(good fit)* und Überanpassung *(overfit)* beim Modelltraining einer Regression. Die blauen Punkte sind die genauen Werte des Datensatzes und die rote Linie gibt die Näherung durch das trainierte Model an. Quelle: XXXXXXXXXXXXXXX

Von einer Überanpassung ist die Rede, wenn das Modell anfängt sich die Beispiele zu merken, anstelle sich die Werte über gefundene Muster und Zusammenhänge selbst zu erschließen. Bei der Vorhersage mit dem Testdatensatz ist das Modell nicht mehr flexibel genug, um eine gute Vorhersage zu treffen X10X. Das bedeutet, dass die Testgenauigkeit sinkt und dadurch die Differenz zwischen Trainings- und Testgenauigkeit signifikant größer wird, was ein Indiz für eine Überanpassung sein kann. Hingegen liegt eine Unteranpassung vor, wenn das Modell unfähig ist, ein Muster in den Trainingsdaten zu erkennen, um daraus einen Trend ableiten zu können. Zu Unteranpassung kommt es, wenn zu wenige Trainingsdaten vorhanden sind, oder die Komplexität der zu erledigenden Aufgabe, die des Modells wesentlich überschreitet. Das Ziel des Trainings ist es, eine Anpassung zu erreichen, die genau zwischen der einer Überanpassung und der einer Unteranpassung liegt.

**Bewertung des Modells:** Zur Bewertung wird das Modell mit gelabelten Daten des Testdatensatzes getestet, welche bei der Datenerfassung von dem gesamten Datensatz abgetrennt wurden. Als Ergebnis erhält man, mehrere Maßzahlen, welche die generelle Leistung des Modells beschreiben (siehe. Statistische Parameter). Zusätzlich können mithilfe einer Validierung die übergeordneten Parameter zur Steuerung des Models *(Hyperparameter)* überprüft und gegebenenfalls angepasst werden. Die Validierung kann mit einem Datensatz gemacht werden, welcher wie der Testdatensatz, dem Trainingsdatensatz vorenthalten wurde (Houldout-Methode). Das Aufteilen auf diese Teildatensätze bringt oftmals das Problem mit sich, dass der Trainingsdatensatz zu klein wird. Um eine Unteranpassung zu vermeiden, kann die Validierung auch mit einer Kreuzvalidierung erfolgen (siehe. 10-fache Kreuzvalidierung).

**Vorhersage:**  Das Modell wird an unbekannten Daten angewandt.

Im Folgenden wird die Funktionsweise der beiden Modelle Kern-Methoden und neuronale Netze näher erläutert.

#### Kern‑Methoden

Mithilfe von Kern‑Methoden (engl. Kernel‑Methods) lassen sich nicht linear klassifizierbare Daten klassifizieren. Für das Verständnis von Kern‑Methoden ist es aber ratsam erst die lineare Klassifizierung zu betrachten.

Bei einer **linearen Klassifizierung** ist das Ziel eine Hyperebene zu finden, welche die Datenpunkte eines Datensatzes in zwei Klassen einteilt (binäre Klassifikation). Im zweidimensionalen ist diese Hyperebene eine Gerade (siehe Abbildung 5). Die optimale Hyperebene wird mit der Funktion beschrieben. Später kann anhand des Vorzeichens der Funktion bei Einsetzung eines zu klassifizierenden Datenpunktes , die Klasse bestimmt werden. Um die optimale Hyperebene zu finden, muss der Rand (Margin) um die Klassengrenzen möglichst breit werden (engl. *large‑margin‑classification*) . Um die optimale Hyperebene zu finden, muss der R befindet sich in der Mitte zwischen den beiden Stützvektoren, welche



Abbildung 5: SVM ...........

Mithilfe des Regulierungsparameters wird die Maximierung des Margin und des Loss Wertes reguliert.

https://www.youtube.com/watch?v=efR1C6CvhmE

Die Support Vector Machine ist ein mathematisches Modell zur Klassifizierung von Objekten, welches von russe…

**Lineare Klassifizierung**

* hard margin verzeiht keine outlieres
* soft verzeiht outlieres 🡺 besser

**nicht -lineare Klassifizierung (inkl, kernel-trick) also daten im kreis**

**daten in höhere dim übertregen, dann wieder trennlinie finden**

Trennung mithilfe einer hyperebene(ab 4 dim davor immer eine dim weniger al daten)

Ziel ist die zuordnung eines neuen obkjektes in eine klasse die sich aus en vorhandenn daten ergeben hat.  
Dazu wird eine optimale Hyperebene bestimmt, anhand dieser später neue Objekte den beiden Klassen zugeordnet werden.

finden einer trennlinie oder trennfläche mit dem größten Abstand zur den nächstliegenden Daten punkten

Bild mit trennlinie und datenpunkrten (linear)

-trennlinie dass 2 daten gut getrennt werden (nur anhand der geraden datenpunkt den beiden klassen zuordnen(vorzeichen)

-wenn zu lange trainiert trenngerade overfitted(nicht mehr perfekt generalisiert) 🡺 mit outlayern

Aber meist nicht linear 🡺 kernel trick = in 3 d übertragen dann da die funbktion finden

Gucken inwiefern ich kernel erkären muss also inwiefern die apg und

#### Neuronale Netze

Erst allgemein 🡺 biologisches vorbild…

Dann dl

https://data-science-blog.com/blog/2017/12/20/maschinelles-lernen-klassifikation-vs-regression/

https://developers.google.com/machine-learning/crash-course/introduction-to-neural-networks/anatomy

Phillip folien [..\deep\_learning\DNN 05 Machine Learning Basics.pdf](file:///C:\Users\Manuel%20Dorer\Desktop\BA_win\deep_learning\DNN%2005%20Machine%20Learning%20Basics.pdf)

„Grundlage neuronale netze „

https://machinelearningmastery.com/what-is-deep-learning/

Tief = viele hiden units (tausende von neuronen)

Input ….hidden(ganz viele)…..output schicht

🡺neuronale netze mit abbildung maybe aus frauenhofer

Selnbständiges lernen und vorhersagen

„Unter Deep Learning versteht man heute neuronalen Netze, die über mehr als einen Hidden Layer verfügen.“

https://www.statworx.com/de/blog/deep-learning-teil-1-einfuehrung/

## ~~KI in der Pharmazeutischen Bioinformatik~~ bessere überschrift vlt datamining

Im datamining kommen Algorithmen des ML zu Einsatz

Drug screening erkären 🡺 mit abb 🡺 basis dafür sind daten

Auf data science und big data zu sprewchen kommen

„sinngemäß schon sehr viele daten erarbeitet aber nicht kompakt“

https://www.sanofi.de/de/sanofi-in-deutschland/news-storys/sanofi-gesundheitstrend-forschung-entwicklung-arzneimittel

<https://www.merckgroup.com/en/research/science-space/envisioning-tomorrow/precision-medicine/generativeai.html>

🡺 große datanbänke gebraucht 🡺 problem dass es diese daten zwar gibt aber nicht schön

Immer mehr literatur... 🡺 nlp

pharma kommen das diese auch (für wirkstoiff finden zb wichtig ist.)

einmal direkt stoff finden

Dass schon viel geforscht wurde (zig millionen paper  unmengen von informatoinen und arbeit) aber unmöglich diese alle durchzulesen  textmining

🡺 probelom in zukunft eher gute datensätze zu haben um damit zu arbeiten

Auf den begriff cpi ppi textmining kommen als anwendung für

## Überblick über State of the Art von “ cpi ppi textmining“

Zeitstrahl inklusive freiburg

- PPI CPI

Freiburg erklären

- aber nur 0 1 ohne weitere klassifizierung

Am ende zu biobert

## Zielsetzung der Arbeit

1) Daten in bracuhbare Form zu überführen

2) vergleichen zu dem paper (freiburg) wie gut sich das problem mit DL lösen lässt 🡺 10cv val

3) abhängigkeit von der größe des training-datensatzes zu bestimmen

Bezug zu paper ganz am ende (freiburg)

# Motivation der Arbeit

Allgemeine entwicklung von ML krass…

Wie in ba „david Baum“

# Material und Methoden

## Systemspecs

Als tabelle

(als envirement)

Tensorflow vs pytorch

## Installation

## Datensätze

Allgemein mit pipelin wie in git

Scripte teilweise erkären

## Statistische Parameter zur Bewertung

https://www.jeremyjordan.me/evaluating-a-machine-learning-model/

Siehe [..\ProceedingsBCVI\_v2\_seite 140.pdf](file:///C:\Users\Manuel%20Dorer\Desktop\BA_win\ProceedingsBCVI_v2_seite%20140.pdf)

Oder doktorarbeit von freidok s 39

## 10-fache Kreuzvalidierung

https://link.springer.com/chapter/10.1007/978-3-319-47217-1\_9

# Ergebnisse

* Gleicher split wie in fr paper zum dir vergleich
* Validation zum Sagen das Hyperparameter okay waren aber auch zum vergleich nutzen
* Veränderung der DS größe um

# Diskussion und Ausblick

Overfitting underfitting?

Ml paper generell fig 1b performance kernek´l methiden schlechter als dl

[einleitung\einleitung\_paper\ML\_paper\_generell.pdf](file:///C:\Users\Manuel%20Dorer\Desktop\BA_win\schreiben\einleitung\einleitung_paper\ML_paper_generell.pdf)

hyperparamter optimierung noch was rausholen

pytoech vs tensorflow (sollte eig nix machen aber vlt überprüfen)

# Zusammenfassung

# Literaturverzeichnis