

Feature Extraktion aus Aerosol-Bildern

Bachelor-Thesis im Studiengang
Elektrotechnik und Informationstechnologie

Modul	TA.BA_BAA+E.H2001
Semester	HS20
Student	Lars Gisler
Dozent	Prof. Dr. Klaus Zahn
Experte	Dr. Jürg Stettbacher
Industriepartner	Swisens AG, Reto Abt und Yanick Zeder
Ort, Datum	Erstfeld, 3. Januar 2021

Inhaltsverzeichnis

1	Einleitung	1
1.1	Swisens AG und ihr Messsystem	1
1.2	Holografie des Swisens Poleno	2
1.3	Feature-Extraktion aus den Holografiebildern	3
1.4	Unzufriedenheiten, Aufgabenstellungen und Ausgangslagen	3
2	Theorieteil	5
2.1	Pollenidentifikation	5
2.1.1	Apertur und Oberflächenbeschaffenheit	5
2.1.2	Polleneinheit	6
2.1.3	Polarität und Form	6
2.2	Bildverarbeitung mit Python	8
2.2.1	FE bisher	8
2.2.2	Wechsel auf OpenCV	10
2.3	Numpy und Pandas	11
2.3.1	Aufbau Dataframe	11
2.4	PCA	12
2.5	SVM	14
2.6	Validierung Klassifikation	15
2.6.1	Confusion Matrix	15
2.6.2	Statistische Gütekriterien der Klassifikation	15
2.6.3	Abhängigkeit des Schwellwertes	16
3	Aufbereitung Daten	17
3.1	Untersuchung Datenqualität und Erstellung Trainings- und Testdatenset	17
3.1.1	Vorgehen	17
3.1.2	Boxplot	18
3.1.3	Einblick in die Bilder und dazugehörigen Features	19
3.1.4	Fazit	19
3.1.5	Erstellung Trainings- und Testdatenset	19
3.2	Genauere Analyse der Daten	20
3.2.1	Vorgehen	20
3.2.2	Kontrolle Bildverarbeitung und erste 3D-Form-Schätzung	21
3.2.3	Neue Definition von Image 0 und Image 1	22
3.2.4	Variation innerhalb Art, Abhängigkeit, Mittelwerte	23
3.2.5	Untersuchung der Orientierung	25
3.2.6	Fazit	25

3.3	Feature-Reduktion mittels PCA.....	27
3.3.1	Vorgehen.....	27
3.3.2	Kovarianzmatrix von X.....	28
3.3.3	Transformationsmatrix P.....	29
3.3.4	Zeilenreduktion der Transformationsmatrix P.....	30
3.3.5	Kovarianzmatrix von Y.....	31
3.3.6	Visualisierung mit PCA.....	32
3.3.7	Fazit PCA.....	32
3.4	Zusammenfassung.....	32
4	Klassifikation der Gattungen.....	33
4.1	Vorgehen.....	33
4.2	Ergebnisse und anschließende Diskussion.....	34
4.2.1	Ranglisten Hyperparameter tuning.....	34
4.2.2	Confusion Matrix und Klassifikationsbericht.....	35
4.3	Abhängigkeit des Scores anhand der Anzahl PCA Features.....	38
4.4	Fazit.....	39
5	Automatische Aussortierung.....	40
5.1	Ablauf Programm.....	40
5.2	Ergebnisse und anschließende Diskussion.....	41
5.2.1	Histogramme der Wahrscheinlichkeitsschätzung.....	41
5.2.2	ROC-Kurven.....	43
5.3	Fazit.....	44
6	Schluss.....	45
6.1	Fazit.....	45
6.2	Ausblick.....	46
6.2.1	Betreff Klassifikation der Gattungen.....	46
6.2.2	Automatische Aussortierung.....	46
6.2.3	Feature-Extraktion.....	46
6.3	Schlusswort.....	47
	Abbildungsverzeichnis.....	48
	Tabellenverzeichnis.....	49
	Literaturverzeichnis.....	50
	Anhang.....	51
	Grafiken automatische Aussortierung.....	51
	Histogramme der Wahrscheinlichkeitsschätzung Trainingsdatenset.....	51
	Histogramme der Wahrscheinlichkeitsschätzung Testdatenset.....	53
	ROC-Kurven.....	55
	Elektronischer Anhang.....	57

1 Einleitung

Dieses Kapitel soll eine Einführung ins Thema der Arbeit geben. Es schildert die Ausgangslage und stellt die bisherigen Problemstellungen und die damit verbundenen Aufgabenstellungen vor.

In einem ersten Teil wird kurz der Industriepartner mit seinem entwickelten Messsystem vorgestellt. Anschliessend wird auf die Holografie tiefer eingegangen. Das ist jener Teil des Messsystems, mit welchem sich die Arbeit auseinandersetzt. Die Arbeit befasst sich zwar nicht mit der Holografie selbst, sondern mit dem Output davon, nämlich mit den aufgenommenen Bildern und deren Verarbeitung. Deswegen wird anschliessend noch grob die bisherige Feature-Extraktion vorgestellt. Der abschliessende Teil des Kapitels stellt dann die bisherigen Unzufriedenheiten und die daraus resultierenden Aufgabenstellungen für die Arbeit vor.

1.1 Swisens AG und ihr Messsystem

Die Swisens AG, ein Start-up-Unternehmen der Hochschule Luzern, hat es sich zum Auftrag gemacht, mehr über die Zusammensetzung der Luft zu erfahren. Ihr entwickeltes Messsystem, der «Swisens Poleno», ermöglicht Echtzeitmessungen von Aerosolpartikeln. Die Partikel können erkannt, gezählt und schliesslich bemessen werden. Dies ermöglicht eine anschliessende Klassifizierung der Aerosolpartikel. Speziell von Interesse sind dabei Pollen, von welchen in der Arbeit hauptsächlich die Rede sein wird.

Der Swisens Poleno beinhaltet drei unterschiedliche Messmethoden, welche nacheinander angewendet werden. Die Partikel werden dabei freischwebend im Luftkanal gemessen. Bei der ersten Messmethodik handelt es sich um eine Holografie. Anschliessend erfolgen eine Fluoreszenzlebensdauer- und Spektrumsmessung. Der letzte Teil beinhaltet eine polarisierte, zeitaufgelöste Streulichtmessung. Die Abbildung 1 zeigt den schematischen Aufbau des Swisens Poleno mit den drei unterschiedlichen Messmethoden. Die Arbeit befasst sich mit einem Teil der Holografie, auf welche im nächsten Abschnitt genauer eingegangen wird. (Swisens, 2020)

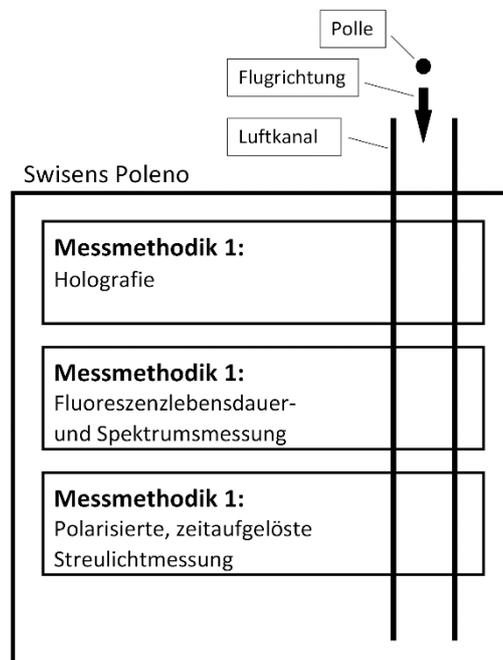


Abbildung 1: Schematischer Aufbau Swisens Poleno

1.2 Holografie des Swisens Poleno

Bei der Holografie werden zwei um 90° verdrehte Bilder eines Partikels aufgenommen. Die Abbildung 2 zeigt den Aufbau der Holografie im Poleno. Sie zeigt gut auf, wie die zwei um 90° verdrehten Aufnahmen aussehen. Die beiden Aufnahmen erfolgen ziemlich schnell aufeinander. Es ist somit mit keiner Rotation des Teilchens zwischen Aufnahme 1 und Aufnahme 2 zu rechnen.

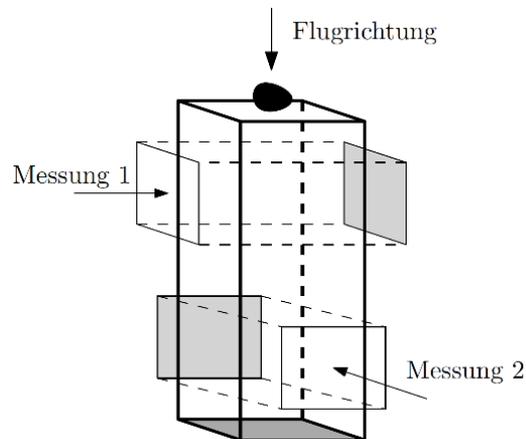


Abbildung 2: Messaufbau Holografie (Bächler, 2017)

Die Gründe, warum die Holografie für die Aufnahme der Bilder angewendet wird und nicht die herkömmliche Technik der Bildaufnahme, sind folgende: Zum einen handelt es sich bei den aufzunehmenden Objekten um Teile im Mikrometerbereich, also um sehr kleine Zielobjekte. Zum anderen muss die Aufnahme, da das Objekt freischwebend im Luftkanal aufgenommen wird, sehr schnell erzeugt werden. Beide dieser Anforderungen können mit der Technik der Holografie erfüllt werden.

Die Holografie funktioniert aufgrund des Wellencharakters des Lichts. Ein Referenzlicht wird dabei auf ein Objekt gerichtet. Die Referenzwellen, welche ungehindert am Partikel vorbeifliegen, und die Wellen, die durch die Beugung des Lichts am Partikel entstehen, erzeugen ein Interferenzmuster auf dem Sensor. Eine Rekonstruktion, welche rein mathematisch vorgenommen wird, kann aus dem Interferenzmuster das Bild der Polle rekonstruieren. Genauer gesagt handelt es sich bei dem rekonstruierten Bild um die Silhouette der Polle. Für die Rekonstruktion des Bildes ist die korrekte Distanz vom Partikel zum Sensor (siehe Abb. 3) von immenser Wichtigkeit. Wird bei der mathematischen Rekonstruktion nicht die korrekte Distanz verwendet, führt dies zu einem unscharfen Bild der Polle. Ein unscharfes Bild führt bei der anschließenden Bildverarbeitung zu ungenaueren Ergebnissen. Vorherige Arbeiten haben sich intensiv mit dem Finden dieser optimalen Distanz auseinandergesetzt. Diese Arbeit befasst sich nun mit dem Ergebnis daraus, mit den rekonstruierten Bildern und deren durch Bildverarbeitung extrahierten Merkmalen. Im Folgenden wird darum noch tiefer auf diesen Teil eingegangen. (Gaussianer, 2020)

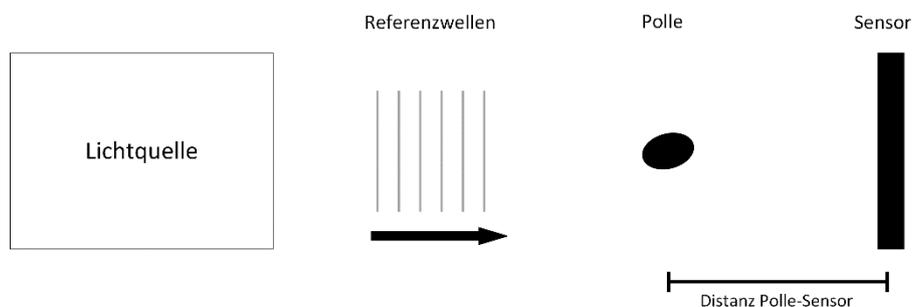


Abbildung 3: Holografie-Funktionsweise

1.3 Feature-Extraktion aus den Holografiebildern

Aus den zwei rekonstruierten Bildern können nun Merkmale mittels Bildverarbeitung extrahiert werden. In einem ersten Teil wird jedoch zuerst das rekonstruierte Bild zugeschnitten. Nur ein 200 * 200 Pixel grosser Ausschnitt, mit idealerweise der Polle im Zentrum des Bildes, wird verarbeitet. In der Abbildung 4 ist ein Beispiel der zwei Ausschnitte ersichtlich. In der Arbeit werden diese Ausschnitte folglich jeweils mit Image 0 und Image 1 bezeichnet.

Es handelt sich dabei um ein Grauwert-Bild mit einer Bit-Tiefe von 16 Bit. Die Feature-Extraktion arbeitet jedoch nur mit einer 8-Bit-Tiefe. Ein Pixel hat dabei eine Abmessung von 0.54 µm.

Beim Output handelt es sich um die extrahierten Features, welche in einem JSON-File abgespeichert werden. Die Tabelle beinhaltet die unterschiedlichen Properties mit einer kurzen Beschreibung. Die Properties aus Image 0 und Image 1 werden dabei separat in der Datenstruktur abgespeichert. Die Funktionsweise der Feature-Extraktion wird im Theorieteil vorgestellt. Hier ist nur kurz der Input und Output davon vorgestellt worden. (Swisens, 2020)

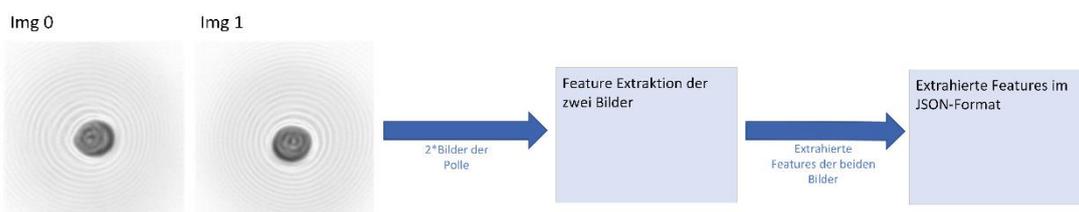


Abbildung 4: Feature-Extraktion

Properties	Einheit	Erklärung
Area	Pixelwerte	Fläche der Polle
Solidity	Zahl	Verhältnis der Fläche der Polle zu ihrer konvexen Hülle
Major Axis	Pixelwerte	Hauptachse einer gefitteten Ellipse der Polle
Minor Axis	Pixelwerte	Nebenachse einer gefitteten Ellipse der Polle
Perimeter	Pixelwerte	Umfang der Polle
Eccentricity	Zahl	Formbeschreibung der gefitteten Ellipse, wird mit Haupt- und Nebenachse berechnet 0 = Kreis, je näher an 1, desto stärker elliptisch
Max Intensity	Zahl von 0 - 1	Maximaler Grauwert in der Fläche der Polle
Min Intensity	Zahl von 0 - 1	Minimaler Grauwert in der Fläche der Polle
Mean Intensity	Zahl von 0 - 1	Mittlerer Grauwert in der Fläche der Polle

Tabelle 1: Extrahierte Features

1.4 Unzufriedenheiten, Aufgabenstellungen und Ausgangslagen

Die bisherigen Unzufriedenheiten der Feature-Extraktion und der Klassifikation der Pollengattungen sind die folgenden:

- Die extrahierten Features stimmen nicht mit den in der Pollenbiologie üblichen Beschreibungen überein.
- Es handelt sich noch nicht um eine einheitliche Beschreibung einer Polle, sondern um zwei separate; von Image 0 und Image 1.
- Bei den extrahierten Features handelt es sich um digitale, nicht physikalische Einheiten.
- Die Feature-Extraktion weist bei kleineren Partikeln noch Schwierigkeiten auf.
- Es gibt noch keine Feature-Extraktion, die speziell für die Bemessung von Sporen ausgerichtet ist.
- Die bisherige Klassifikation der Pollengattungen basiert auf einem Convolutional Neural Network (CNN), welches direkt auf die Bilder der Pollen angewendet wird, und ist somit mehr oder weniger eine Blackbox.

Zudem hat sich im Laufe der Arbeit gezeigt, dass die Datensets teils noch fehlerbehaftete Daten (also schlechte Bilder) beinhalten.

Daraus resultieren folgende Aufgaben:

- 1) Mit einer Recherche über die Pollenbiologie sollen die für die Beschreibung einer Polle üblichen Merkmale gefunden werden. Und zwar jene Merkmale, welche auch mittels einer Bildverarbeitung von Image 0 und 1 herausgeholt werden können.
- 2) Mit einer Klassifikation der Pollengattung mit den extrahierten Merkmalen soll eine Alternative zur bisherigen Klassifikation erarbeitet werden.
- 3) Es soll eine automatische Aussortierung (Säuberung) von Datensets erstellt werden.
- 4) Anpassung der FE, damit obige Punkte behoben werden (Anpassung an neue Merkmaliste, einheitliche Beschreibung einer Polle, physikalische Einheiten berechnen; beinhaltet zudem 3D-Form-Schätzung).

Es resultieren folgende Zusatzanforderungen:

- Feature-Extraktion bei kleinen Partikeln verbessern
- Feature-Extraktion für Bemessung von Sporen anpassen
- Performance-Verbesserungen

Pollengattungen

In der Liste sind die 14 Pollengattungen ersichtlich, zwischen welchen unterschieden wird. Es ist jeweils der lateinische wie auch der deutsche Name angegeben. In der Dokumentation wird folglich ausschliesslich der lateinische Name verwendet.

Index	Name lateinisch	Name deutsch
1	Alnus	Erlen
2	Carpinus	Hainbuche
3	Corylus	Hasel
4	Cryptomeria	Sicheltanne
5	Cupressus	Zypressen
6	Dactylis	Wiesen-Knäuelgras
7	Fagus	Buchen
8	Fraxinus	Eschen
9	Gram	kein deutscher Name
10	Juncaceae	Binsengewächse
11	Populus	Pappeln
12	Quercus	Eichen
13	Taxus	Eiben
14	Ulmus	Ulmen

Tabelle 2: Liste der Pollengattungen

Zur Verfügung stehende Daten

Der Industriepartner stellt von jeder Gattung zwei bis vier Ordner mit ausreichend vielen Daten von Pollen zur Verfügung. Zu den Daten einer Polle gehören jeweils die zwei Ausschnitte der rekonstruierten Aufnahme, also Image 0 und Image 1, sowie ein JSON-File mit Informationen zur Messung allgemein und den bisherig extrahierten Features (siehe Feature-Extraktion).

Gearbeitet wird jeweils mit dem ersten Ordner der Gattung. Dieser umfasst schon sehr viele Daten. Es wird davon ausgegangen, dass dieser Ordner für die jeweilige Gattung repräsentativ ist.

2 Theorieteil

Dieses Kapitel stellt die Theorie, welche für die Arbeit relevant ist, kurz und einfach vor. Es wird dabei nicht zu tief ins Detail gegangen.

Der erste Teil befasst sich mit der Pollenidentifikation, welche für die Erstellung der neuen Merkmalliste gebraucht wird. Anschliessend wird kurz auf die Feature-Extraktion eingegangen. Danach hat man genaueres Verständnis, wie man mittels der Bildverarbeitung zu den extrahierten Features kommt, welche bei der Arbeit von grosser Bedeutung sind. Mit Numpy und Pandas werden die Libraries vorgestellt, welche einen einfachen Umgang mit grossen Datensätzen ermöglichen. Zudem wird kurz auf die verwendete Struktur des Panda Dataframe eingegangen und die wichtigsten Operationen kurz erläutert. PCA stellt kurz theoretisch das Werkzeug vor, welches für die Reduktion der Datendimensionalität gebraucht wird. Zudem wird die verwendete Notation vorgestellt. Es wird kurz die Support Vektor Machine (SVM) vorgestellt, welche für die Klassifizierung der Gattungen gebraucht wird. Der letzte Teil befasst sich mit den Grössen und Grafiken, welche für die Validierungen der Klassifikation der Gattungen sowie der automatischen Aussortierung verwendet wurden.

2.1 Pollenidentifikation

Das Buch «Illustrated Pollen Terminology», welches für die Recherche vom Industriepartner vorgeschlagen wurde, teilt die Beschreibung einer Polle in folgende Punkte auf: Apertur (Öffnungen), Polleneinheit, Polarität und Form sowie Oberflächenbeschaffenheit. Diese werden nun kurz erläutert. Apertur und die Oberflächenbeschaffenheit werden dabei zusammen vorgestellt.

2.1.1 Apertur und Oberflächenbeschaffenheit

Bei der Apertur werden die Keimöffnungen und Poren einer Polle beschrieben. Dabei können unterschiedliche Anzahlen, Formen und Konstellationen auftreten. Es ist zu vermerken, dass die Keimöffnungen auch einen Einfluss auf die Form einer Polle nehmen können. Mit verschiedenen Messmethoden kann die Oberflächenbeschaffenheit beschrieben werden. In der Abbildung ist eine Polle mit drei Keimöffnungen und einer punktförmigen Oberflächenbeschaffenheit ersichtlich (Heidemarie Halbritter, 2018).

Die neu erstellte Merkmalliste für die Feature-Extraktion enthält nicht die Beschreibung der Apertur und der Oberflächenbeschaffenheit, da die beiden Umrisse der Pollen für eine solche Angabe nicht ausreichen.



Abbildung 5: Apertur und Oberflächenbeschaffenheit (AutPal, 2020)

2.1.2 Polleneinheit

Mehrere einzelne Pollenkörner können zusammen als Einheit auftreten. Dabei wird diese Einheit bezüglich Anzahl der einzelnen Pollenkörner sowie der Art der Konstellation beschrieben. Die Abbildung 6 zeigt einige mögliche Einheiten auf. Weitere Konstellationen sind im Anhang «Quick Reference Glossary with Illustrations» zu finden (Heidemarie Halbritter, 2018).

Die Swisens AG hat mitgeteilt, dass Einheiten von mehreren einzelnen Pollenkörnern eher selten auftreten. Deswegen wird diese Beschreibungsform nicht in die neue Merkmalliste hineingenommen. Man beschränkt sich nur auf ein einzelnes Pollenkorn. Es ist aber durchaus denkbar, dass dies in einem weiteren Schritt miteinbezogen werden könnte. Dazu ist zu sagen, dass sich nur gewisse einfache Einheiten fürs Beschreiben eignen. Denn wenn sie zu dicht beieinander sind, kann man sie anhand des Umrisses nicht identifizieren.

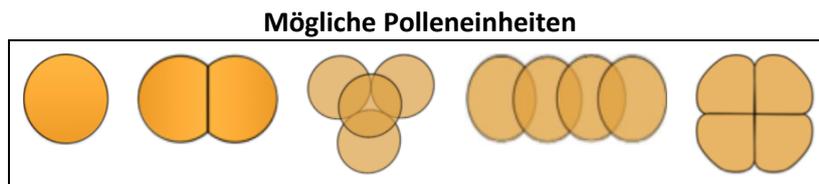


Abbildung 6: Mögliche Polleneinheiten (Huffner)

Grösse

Zur Polleneinheit gehört auch die Angabe der Grösse, welche durch die Feature-Extraktion gut bestimmt werden kann. Die Tabelle zeigt eine gebräuchliche Abstufung der Grösse. Es handelt sich dabei um die längste Achse einer Polle. Weiter kann man die kürzeste und längste Achse sowohl von der Polar- als auch Äquatorialansicht angeben.

very small	<10 μm
small	10 – 25 μm
medium	26 – 50 μm
large	51 – 100 μm
very large	>100 μm

Tabelle 3: Grössenunterteilung Pollen (Heidemarie Halbritter, 2018)

2.1.3 Polarität und Form

Polarität und Form sind für die Feature-Extraktion wohl am meisten von Bedeutung. Diese Merkmale sollten anhand der beiden Umrisse berechnet oder geschätzt werden. Nachfolgend werden die wichtigsten Merkmale kurz vorgestellt.

Polar- und Äquatorialansicht

Eine Polle besitzt jeweils eine polare sowie eine äquatoriale Ansicht. Beide Ansichten können unterschiedliche Formen annehmen. Die Abbildung 7 zeigt jeweils fünf mögliche Formen auf. Weitere sind im Anhang «Quick Reference Glossary with Illustrations» ersichtlich.

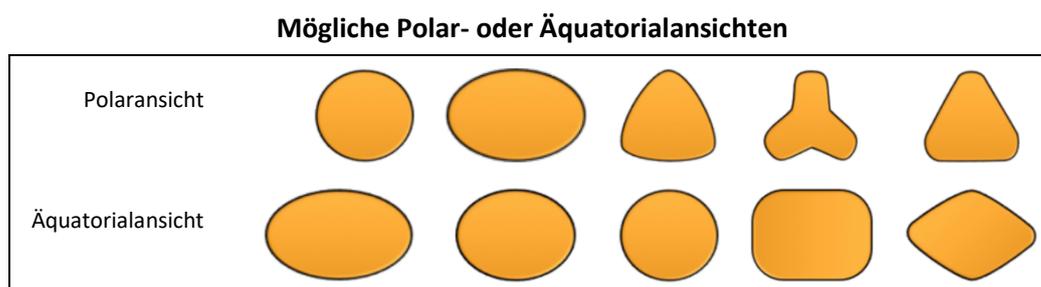


Abbildung 7: Polar- und Äquatorialansichten (Huffner)

PE-Verhältnis und 3D-Form

Bei der 3D-Form ist vor allem das PE-Verhältnis von Bedeutung. Es handelt sich dabei um das Verhältnis der Distanz zwischen den beiden Polen (Durchmesser äquatoriale Ansicht) zum Durchmesser der polaren Ansicht. Die Tabelle zeigt eine Einstufung von verschiedenen Verhältnissen. Bei der 3D-Form kann es anhand der unterschiedlichen Kombination von polarer und äquatorialer Form zu vielen verschiedenen Formen kommen. Die Möglichkeiten sind im Anhang «Illustrated Pollen Terms» ersichtlich.

Oblate		Prolate	
Name	PE-Verhältniss	Name	PE-Verhältnis
Peroblate	< 0.5	Perprolate	> 2
Oblate	0.5 – 0.75	Prolate	1.33 – 2
Suboblate	0.75 – 0.88	Subprolate	1.14 – 1.33
Oblate spheroidal	0.88 – 1	Prolate spheroidal	1 – 1.14
Subspheroidal PE-Verhältnis von 0.75 – 1.33			

Tabelle 4: PE-Verhältnisse (Erdtman, 1986)

Für die Beschreibung am wichtigsten ist das Erkennen einer Polar- und Äquatorialansicht sowie die dazugehörigen Längen der Achsen. Alle weiteren Merkmale wie Form und Verhältnis können dann daraus abgeleitet werden.

2.2 Bildverarbeitung mit Python

Python ist eine beliebte und viel benutzte Programmiersprache für die Bildverarbeitung. Sie besitzt eine Grosszahl von unterschiedlichen Libraries. Einige Libraries implementieren die Bildverarbeitungsfunktionen in C oder C++, wodurch Python auch für Echtzeitanwendungen verwendet werden kann. (Pandey, 2020)

2.2.1 FE bisher

Die bisherige Feature-Extraktion basiert auf der Scikit-image Library. Diese ist, wie so viele andere Libraries auch, auf der Numpy Library aufgebaut. Ein Bild wird somit mit einem Numpy Array dargestellt. Die Library verfügt über eine Grosszahl von Bildverarbeitungsfunktionen. Sie ist einfach anzuwenden und verfügt über eine gute Dokumentation. (scikit-image-development-team, 2020)

Der Ablauf der bisherigen Bildverarbeitungsfunktion sieht wie folgt aus:

Input:

Als Input dient, wie bereits in der Einleitung vorgestellt, ein 200*200 Pixel Grauwert Image mit 16-Bit-Tiefe. Gearbeitet wird jedoch nur mit 8-Bit-Tiefe. Auf dem Bild befindet sich idealerweise immer ein Abbild einer Polle, und diese mit möglichst scharfen Kanten. Zusätzlich wird auch noch eine Datenstruktur mitgegeben, welche beim Return abgefüllt wird.

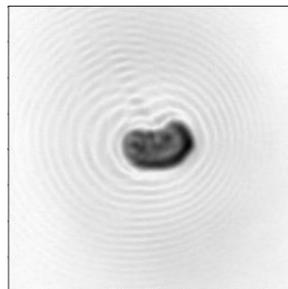


Abbildung 8: Beispielbild Input

Schritt 1: Finden des Thresholds

In einem ersten Schritt wird mit der Otsu-Funktion ein Threshold gefunden. Wie in der Abbildung am Histogramm ersichtlich, berechnet die Otsu-Methode für das Beispielbild einen Threshold von 180.

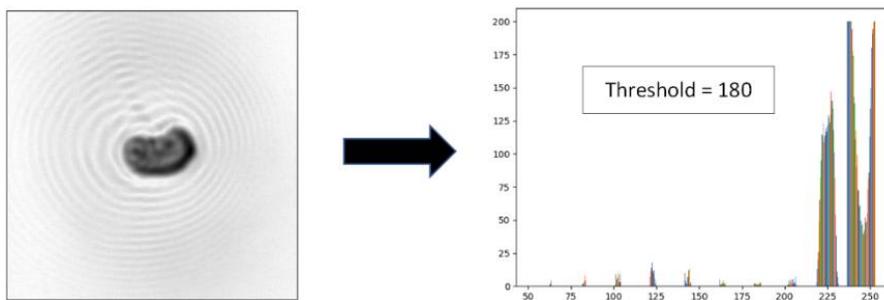


Abbildung 9: Finden des Thresholds

Schritt 2: Erstellung des Binary Image

Mit dem gefundenen Threshold kann nun ein Binary Image erstellt werden.

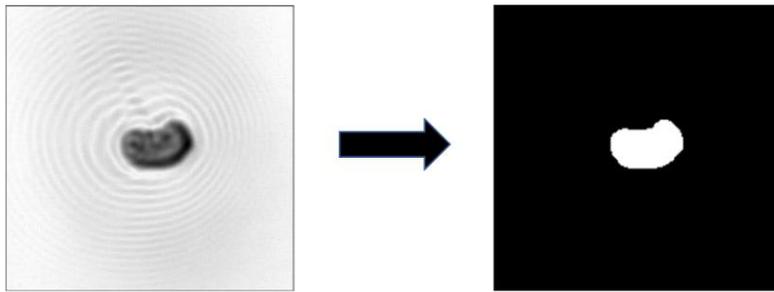


Abbildung 10: Erstellung des Binary-Bildes

Schritt 3: Labeln des Binary Image

In einem weiteren Schritt werden die Cluster auf dem Binary-Bild gelabelt. Da auf dem Beispielbild nur ein Cluster enthalten ist, wird diesem die Nummer 1 zugeteilt.

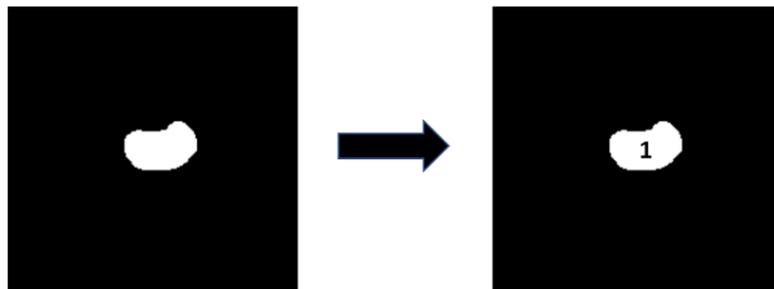


Abbildung 11: Labeln des Binary Image

Schritt 4: Finden des grössten Clusters

Mit einer weiteren Funktion wird dann das grösste Cluster auf dem Bild gefunden. Also jenes Cluster, welches die grösste Fläche besitzt.

Schritt 5: Region Properties berechnen

Anschließend werden die Merkmale des grössten Clusters im originalen Bild berechnet. Es handelt sich dabei um die Merkmale, welche bereits in der Einleitung vorgestellt wurden.



Abbildung 12: Region-Properties-Berechnung

Output

Beim Return werden der mitgegebenen Datenstruktur die einzelnen Properties zugewiesen.

2.2.2 Wechsel auf OpenCV

Die Feature-Extraktion wird neu mit der OpenCV Library realisiert. OpenCV ist die meistverwendete Python Library für Bildverarbeitung. Sie verfügt somit über eine sehr grosse Community, und es finden sich bei Problemen oft schnell einige Lösungen. OpenCV implementiert die meisten Funktionen mit C oder C++ und ist somit geeignet für Echtzeitanwendungen. (OpenCV-team, 2020)

Mit dem Wechsel auf OpenCV werden einerseits Performance-Verbesserungen und andererseits mehr Möglichkeiten für den Ausbau der Feature-Extraktion erhofft.

Berechnung der bisherigen Features

In einem ersten Schritt sind mit OpenCV die bisherigen Features extrahiert worden.

Der erste Teil des Quellcodes ähnelt dabei sehr stark dem bisherigen. Nur bei der Berechnung der Properties weicht der neue Code etwas vom bisherigen ab. In der Scikit-image Library gibt es eine Funktion, welche alle Properties direkt berechnet, und in OpenCV müssen die Properties einzeln anhand der Kontur berechnet werden. Dies ist ein wenig aufwendiger, hat aber den Vorteil, dass nur das berechnet wird, was auch gebraucht wird.

Neu hinzugenommene Features

Die Feature-Extraktion wurde zusätzlich um einige Features erweitert. Hinzugekommen ist eine Information der Orientierung (siehe Abbildung 13, links). Es handelt sich dabei um den Winkel der Hauptachse der gefitteten Ellipse zur vertikalen Achse. Der Winkel wird in Grad angegeben. Dieses Feature wird dann wahrscheinlich für die 3D-Form-Schätzung wichtig sein.

Zudem werden die Maximalpunkte der Polle auf allen vier Seiten berechnet (siehe Bild rechts). Daraus können dann die maximale Höhendifferenz und die maximale Differenz in der Horizontalen berechnet werden. Die maximale Höhendifferenz kann vor allem für die Begutachtung der beiden Bilder genutzt werden. Die Differenz sollte bei beiden Bildern gleich gross sein. Ist dies nicht so, kann man davon ausgehen, dass irgendetwas nicht korrekt funktioniert hat.

Neu wird auch der Schwerpunkt der gefitteten Ellipse sowie der Polle extrahiert.

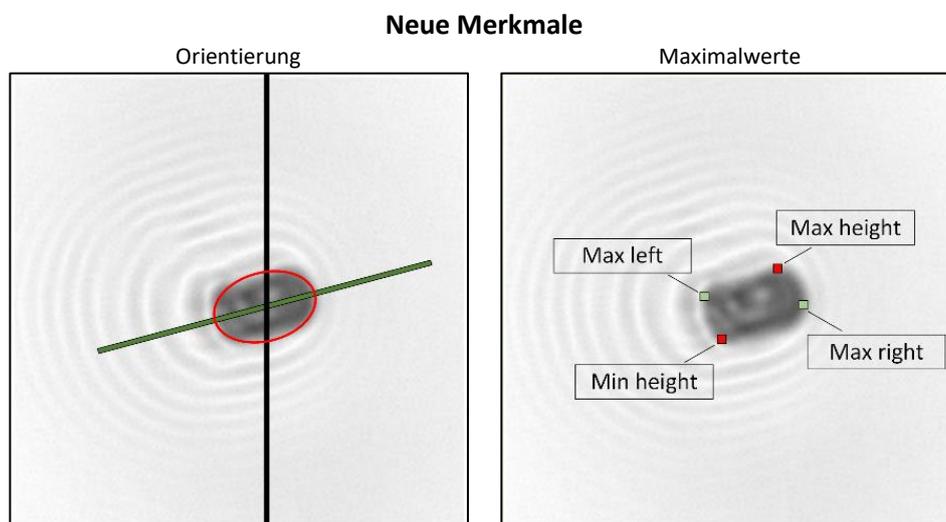


Abbildung 13: Neue Merkmale

2.3 Numpy und Pandas

Die Numpy und die Pandas Library sind beim Arbeiten mit Python von grösster Wichtigkeit. Sie vereinfachen den Umgang mit den grösseren Daten enorm. Viele andere Libraries wie z. B. Visualisierungs-Libraries oder Machine Learning Libraries sind oft auf diesen aufgebaut. (Hossen, 2020)

Numpy ermöglicht die Vektor- und Matrizenrechnung. Sie inkludiert viele mathematische Operationen. Ausserdem ist die Syntax sehr ähnlich wie in Matlab. Dies vereinfacht einen Wechsel sehr.

Pandas erweitert dann das Ganze noch mit dem Pandas Dataframe. Das ist eine Datenstruktur, die einer Tabelle sehr ähnelt. Man kann den einzelnen Kolonnen und Spalten Namen geben. Das Pandas Dataframe ermöglicht einen einfachen Import und Export der Daten als CSV-Files.

2.3.1 Aufbau Dataframe

In der Tabelle 5 ist ersichtlich, wie das Pandas Dataframe für das Trainings- und Testdatenset in der Arbeit aufgebaut ist. Eine Zeile beinhaltet die Informationen von den zwei Bildern aus der Holografie.

Gelb markiert sind die Informationen, welche für das File-Handling gebraucht werden. Dazu gehört der Datenpfad (dir) zu den Daten sowie die Namen von Image 0 und 1 und des jeweiligen JSON-Files.

Blau markiert sind Labels (_lab) sowie direkt ein Platzhalter für die Schätzung (_pred). Zum einen für die Qualität (quality), welche bei der automatischen Aussortierung gebraucht wird (siehe Kap. 5). Die Codierung der Qualität ist in der Tabelle 6 darunter ersichtlich. Zum anderen für die Gattung der Polle (genus), welche hauptsächlich für die Klassifizierung der Gattungen gebraucht wird. Es handelt sich dabei um die in der Einleitung vorgestellten 14 Gattungen.

Im grünen Teil sind dann die extrahierten Features aus Image 0 und 1. Die Nummern nach den Feature-Namen zeigen an, zu welchem Image das extrahierte Feature gehört.

Aufbau Dataframe

	File-Handling				Labeling und Prediction				Features																		
Index/Sample	dir	name_event	name_im0	name_im1	quality_lab	quality_pred	genus_lab	genus_pred	solidity_0	solidity_1	area_0	area_1	minorAxis_0	minorAxis_1	majorAxis_0	majorAxis_1	perimeter_0	perimeter_1	maxIntensity_0	maxIntensity_1	minIntensity_0	minIntensity_1	meanIntensity_0	meanIntensity_1	eccentricity_0	eccentricity_1	
1																											
2																											
3																											
...																											

Tabelle 5: Aufbau Dataframe

Mit Operationen können aus dem Dataframe nun einfach die gewünschten Informationen herausgeholt werden, welche für die einzelnen Arbeiten erforderlich sind. Wie zum Beispiel das Extrahieren der Datenmatrix (nur grüner Teil), welche dann direkt in ein Numpy Array umgewandelt werden kann. Oder das Extrahieren eines einzelnen Vektors, wie zum Beispiel die Spalte genus_lab, für eine Klassifizierungsaufgabe.

Wichtig ist auch das Extrahieren von Daten, welche eine bestimmte Bedingung erfüllen. Zum Beispiel die Daten einer bestimmten Gattung oder mit einer guten Qualität. Auch dies konnte einfach realisiert werden.

Dummy-Variable	Bedeutung
1	Gut
-1	Schlecht
0	Undefiniert

Tabelle 6: Codierung Qualität

2.4 PCA

PCA (Principal Component Analysis / deutsch: Hauptkomponentenanalyse) ist ein mathematisches Verfahren, mit welchem die Dimension eines Datensatzes reduziert werden kann. Es handelt sich dabei um einen unsupervised Algorithmus. Er eignet sich für die Reduktion von Features. Er ist auch beliebt für Visualisierungen von mehrdimensionalen Daten. (Shlens, 2003)

Bevor auf die eigentliche Erklärung von PCA eingegangen wird, erfolgt eine kurze Erläuterung der allgemeinen Notation, der Standardisierung und der Kovarianzmatrix.

Allgemeine Notation

Die Tabelle 7 beinhaltet die gesamte Notation, welche im Folgenden gebraucht wird. Es muss aufgepasst werden, dass in den unteren Rechnungen jeweils mit dem Transponierten von P gerechnet wird und somit m2 und n2 jeweils vertauscht werden müssen.

X	Datenmatrix	m1 = Anzahl Samples n1 = Anzahl originale Features
P	Transformationsmatrix	m2 = Anzahl PCA Features n2 = Anzahl originale Features
Y	Transformierte Datenmatrix	m3 = Anzahl Samples n3 = Anzahl PCA-Features

Tabelle 7: Notation PCA

Standardisierung

Bevor der PCA-Algorithmus angewendet wird, ist es wichtig, die Datenmatrix zu standardisieren, sodass der Erwartungswert jeweils null ergibt und die Varianz eins. Für diese Transformation müssen zuerst die Mittelwerte μ und Standardabweichungen σ der einzelnen Features berechnet werden. Die Formel für die Standardisierung sieht wie folgt aus: (Myrianthous, 2020)

$$X = \frac{X_{\text{unscaled}} - \mu}{\sigma}$$

Kovarianzmatrix

Es wird nur kurz erläutert, was anhand der Kovarianzmatrix der Datenmatrix ersichtlich ist. Bei der Diagonalen handelt es sich um die Varianz der Features. Die Varianz sagt etwas über den Informationsgehalt in einem Feature. Bei den nicht diagonalen Einträgen handelt es sich jeweils um die Kovarianz zweier Features. Ist diese ungleich null, deutet dies auf Redundanz zwischen den beiden Features. Die Kovarianzmatrix ist symmetrisch. (Wikipedia, 2020)

PCA

Bei PCA handelt es sich um eine Basistransformation eines Datensatzes. P rotiert und stretcht den Datensatz von X zu Y. Bei den Zeilen von P handelt es sich um die neuen Basisvektoren vom Datensatz Y. Die Zeilen werden Principal Components genannt.

Ziel ist es, ein P zu finden, welches Folgendes erfüllt:

- Die nicht diagonalen Einträge der Kovarianzmatrix von Y sind gleich null. Was bedeutet, dass keine Redundanz zwischen den neuen PCA-Features mehr vorhanden ist.
- Die Diagonalwerte der Kovarianzmatrix von Y, also die Varianz der neuen PCA-Features, werden nach der Grösse geordnet. So kann später die Reduktion anhand des Informationsgehalts sehr einfach vorgenommen werden.

$$\begin{matrix} m1 \\ \left[\begin{array}{c} \mathbf{X} \\ \hline \end{array} \right] \\ n1 \end{matrix} * \begin{matrix} m2 \\ \left[\begin{array}{c} \mathbf{P}^T \\ \hline \end{array} \right] \\ n2 \end{matrix} = \begin{matrix} m3 \\ \left[\begin{array}{c} \mathbf{Y} \\ \hline \end{array} \right] \\ n3 \end{matrix}$$

P wird über die Eigenvektoren und Eigenwerte der Kovarianzmatrix von X berechnet. Bei den Zeilen von P handelt es sich um die Eigenvektoren von der Kovarianzmatrix von X , welche nach der Grösse der Eigenwerte geordnet sind.

Die neuen PCA Features sind nichts anderes als lineare Abbildungen der bisherigen Features. Anhand der Werte der Eigenvektoren ist die Gewichtung des alten Features für das neue PCA Feature ersichtlich.

Reduktion der Datendimensionalität

Wird n_2 verkleinert, also Zeilen aus P werden herausgenommen (von unten nach oben), wird die Dimensionalität von Y reduziert.

Die Eigenwerte entsprechen der Varianz der neuen PCA-Features. Ist diese gleich null, kann dieses Feature verworfen werden, ohne einen Informationsverlust zu generieren.

Anschliessend können diese Features eliminiert werden, welche nur einen sehr kleinen Anteil des Gesamtinformationsgehaltes ausmachen. Dabei weiss man genau, wie viel Prozent an Information man verliert.

Kurz zusammengefasst

Die neuen PCA Features sind lineare Kombinationen der bisherigen Features.

Mit der Reduktion von n_2 (Zeilen von P) reduziert man die Datendimension von Y .

Anhand der Eigenwerte ist bekannt, wie viel an Information man verliert.

(Shlens, 2003)

2.5 SVM

Die Support Vector Maschine (SVM) ist ein Klassifikator, mit welchem sowohl binäre Klassifizierungen als auch Mehrklassen-Klassifikationen realisiert werden können. Es handelt sich um einen Large Margin Klassifizierer, was so viel bedeutet, dass versucht wird, die «besten» Grenzen zu finden, nämlich solche, welche den grössten Abstand zu den Datenpunkten haben. Dies ist in der Abbildung 14 gut ersichtlich. Die linke Grafik zeigt einige von vielen möglichen Grenzen, die rechte wird mit einer SVM gefunden. (scikit-learn-authors, 2020)

Eine SVM ist beim Trainieren eher rechenintensiv. Das trainierte Modell beinhaltet nur die sogenannten Support-Vektoren, welche in der Abbildung ausgemalt dargestellt werden, und ist deswegen meist nicht mehr so rechenintensiv.

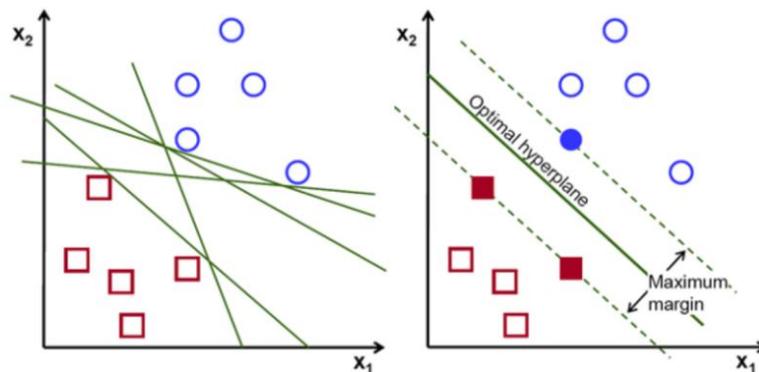


Abbildung 14: SVM (Chakure, 2020)

Kernel-Trick und Hyperparameter

Mit dem Kernel-Trick lassen sich auch nicht linear trennbare Probleme lösen. Je nach Kernel-Typ hat es unterschiedliche Hyperparameter. Es empfiehlt sich, mit Hyperparameter-Tuning den richtigen Kernel-Typ und dazugehörige Parameter zu finden. Dies kann in der Scikit-Learn Library zum Beispiel mit einem Gridsearch und einer Crossvalidation realisiert werden. (scikit-learn-authors, 2020)

Die Tabelle beinhaltet einige möglich Kernel sowie dazugehörige Parameter.

Kernel-Typ	Dazugehörige Hyperparameter
linear	C
rbf (Radial Basis Function)	C, gamma
sigmoid	C, gamma
Poly (polynomial)	S, gamma, degree

Tabelle 8: Kernel-Typen und Parameter

2.6 Validierung Klassifikation

Bei einer Klassifikation kann es zu zwei Typen von korrekten Entscheiden und zu zwei Typen von falschen Entscheiden kommen. True Positive (TP) und True Negative (TN) sind die beiden korrekten Entscheide und False Positive (FP) und False Negative (FN) sind die beiden nicht korrekten Entscheide, welche auch als Error Typ 1 und Error Typ 2 bezeichnet werden. (Haß, 2020)

Die untere Tabelle erläutert diese Typen an einem Beispiel, bei welchem ein Klassifikator entscheiden muss, ob sich auf einem Bild ein Hund befindet oder nicht.

TP	True Positive	Klassifikator entscheidet, dass sich ein Hund auf dem Bild befindet, wo auch ein Hund drauf zu sehen ist
TN	True Negative	Klassifikator entscheidet, dass sich kein Hund auf dem Bild befindet, wo auch kein Hund auf dem Bild zu sehen ist
FP	False Positive (Error Typ 1)	Klassifikator entscheidet, dass sich ein Hund auf dem Bild befindet, obwohl kein Hund auf dem Bild zu sehen ist
FN	False Negative (Error Typ 2)	Klassifikator entscheidet, dass sich kein Hund auf dem Bild befindet, obwohl ein Hund auf dem Bild zu sehen ist

Tabelle 9: Mögliche Ergebnisse bei einer Klassifikation

2.6.1 Confusion Matrix

Die Confusion Matrix (deutsch: Wahrheitsmatrix) ist eine gängige Art, diese unterschiedlichen Typen in einer Tabellenform darzustellen. Häufig handelt es sich bei den Spalten um den wahren Wert (actual class) und bei den Zeilen um den geschätzten Wert (predicted class). Eine Confusion Matrix kann auch bei einer Mehrklassen-Klassifikation erstellt werden. (Haß, 2020)

		Actual class	
		P	N
Predicted class	P	TP	FP
	N	FN	TN

Tabelle 10: Confusion-Matrix-Beispiel

2.6.2 Statistische Gütekriterien der Klassifikation

Mit den Werten TP, TN, FN und FP können viele unterschiedliche Gütekriterien berechnet werden. Im Folgenden werden kurz jene vorgestellt, welche bei der Klassifikation in der Arbeit gebraucht werden. (Haß, 2020)

Precision

Auch positiver Vorhersagewert genannt. Dieser Wert sagt, wie viele von denen, welche als positiv klassifiziert wurden, auch wirklich positiv sind. Die erste Zeile der Matrix ist dabei entscheidend. Die Formel sowie die bedingte Wahrscheinlichkeit sehen wie folgt aus:

$$PPV = \frac{TP}{TP + FP} = P(\text{tatsächlich positiv} \mid \text{positive Klassifikation})$$

Recall (Trefferquote)

Auch Sensitivität genannt. Dieser Wert sagt, wie viele von den wahren Positiven auch als positiv klassifiziert werden. Die erste Spalte der Matrix ist dabei entscheidend. Die Formel sowie die bedingte Wahrscheinlichkeit sehen wie folgt aus:

$$TPR = \frac{TP}{TP + FN} = P(\text{positive Klassifikation} \mid \text{tatsächlich positiv})$$

F1-Score

Der F1-Score kombiniert die Precision und den Recall. Es handelt sich dabei um das gewichtete harmonische Mittel. Die Formel sieht wie folgt aus:

$$F1 = 2 \cdot \frac{PPV \cdot TPR}{PPV + TPR} = \frac{2TP}{2TP + FP + FN}$$

Accuracy

Auch Vertrauenswahrscheinlichkeit oder Treffergenauigkeit genannt. Dieser Wert gibt den Anteil der als korrekt klassifizierten Ergebnisse an. Berechnet wird er mit der Diagonalen der Matrix über der Gesamtzahl der Matrix. Die Formel sowie die bedingte Wahrscheinlichkeit sehen wie folgt aus:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} = P(\text{korrekt klassifiziert})$$

2.6.3 Abhängigkeit des Schwellwertes

Histogramm

Wird die Entscheidung eines Klassifikators anhand einer Wahrscheinlichkeit entschieden, so eignet sich oft auch die Darstellung der beiden Verteilungskurven (wahr positiv (P) rechts und wahr negativ (N) links). Diese können beispielsweise auch mit einem Histogramm dargestellt werden.

Die Horizontale Linie symbolisiert den Schwellwert, welcher nach links oder rechts verschoben werden kann. Diese Darstellung zeigt gut auf, dass die Größen TP, TN, FP und FN von einem Schwellwert abhängig sind. Wird zum Beispiel der Schwellwert nach rechts verschoben, vergrößern sich TN und FN. Zugleich verkleinern sich FP und TP. Je weniger stark die Kurven ineinander liegen, desto besser. Der Schwellwert kann nun je nach Bedürfnissen eingestellt werden. (roc-curves-Wikipedia, 2020)

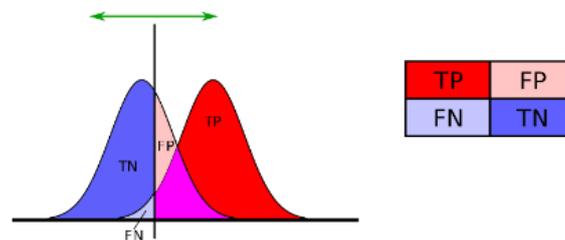


Abbildung 15: Abhängigkeit eines Schwellwertes mittels Verteilung (roc-curves-Wikipedia, 2020)

ROC-Kurve

Ein anderes beliebtes Werkzeug, um die Abhängigkeit des Schwellwertes ersichtlich zu machen, ist die ROC-Kurve. Meistens handelt es sich bei der X-Achse um die False Positive Rate (FPR) und bei der Y-Achse um die True Positive Rate (TPR). Anhand dieser Kurve sind direkt die Raten, welche erreicht werden können, ersichtlich. Je näher die Kurve an den Punkt (0,1) kommt, desto besser. Dies bedeutet eine hohe TPR bei einer geringen FPR. (roc-curves-Wikipedia, 2020)

Die Kurve entsteht mit dem Verschieben des Schwellwertes wie in der obigen Grafik von ganz rechts nach ganz links – von einem Wahrscheinlichkeits-Schwellwert zum Beispiel von 100 % auf 0 %.

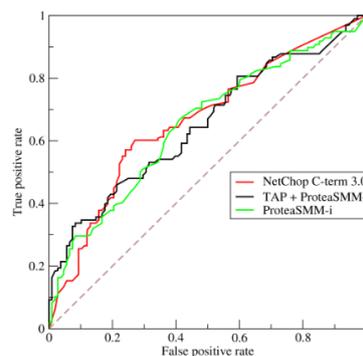


Abbildung 16: ROC-Kurve (roc-curves-Wikipedia, 2020)

3 Aufbereitung Daten

Dieses Kapitel dient dazu, die Daten für eine folgende Klassifizierung der Gattungen und eine automatische Aussortierung vorzubereiten.

In einem ersten Teil soll die Qualität der zur Verfügung stehenden Daten untersucht werden und ein Trainings- sowie ein Testdatenset erstellt werden. Eine zweite tiefere Untersuchung, welche an den guten Daten am Trainingsset durchgeführt wird, soll mehr Einsicht in die Daten geben. Es soll festgestellt werden, welche Features auf welche Weise ins Datenset aufgenommen werden. In einem letzten Teil wird mittels PCA eine Reduktion der Dimension des Datensets vorgenommen.

3.1 Untersuchung Datenqualität und Erstellung Trainings- und Testdatenset

Bei dieser Untersuchung soll ein erster Überblick über die zur Verfügung stehenden Daten geschaffen werden. Fehlerbehaftete Daten und ihre Quellen sollen auffindig gemacht werden. Allenfalls sollen bereits gefundene Besonderheiten der Daten, welche für das weitere Vorgehen nützlich sein könnten, festgehalten werden.

Mit dem erworbenen Wissen soll es möglich sein, einen Trainings- sowie einen Testdatensatz zusammenzustellen, mit welchen dann weitergearbeitet werden kann.

3.1.1 Vorgehen

Es wird mit dem ersten Ordner einer Gattung gearbeitet. Es wird davon ausgegangen, dass dieser repräsentativ ist (siehe Einleitung).

Schritt 1: Boxplot

In einem ersten Teil werden einzelne bisher extrahierte Features eingelesen und in einem Boxplot dargestellt. Die Grafik soll helfen, allfällige Ausreisser zu erkennen, um so einen ersten Überblick über die Qualität der ermittelten Daten zu erhalten. Zudem gibt der Boxplot bereits einen ersten Einblick über die Variation zwischen den Arten und grob über die Variation innerhalb einer Art.

Schritt 2: Einblick in die Bilder und dazugehörigen extrahierten Features

In einem zweiten Teil sollen mit einem Einblick in die Bilder die Fehlerquellen und ihre Auswirkungen gefunden werden.

3.1.2 Boxplot

Es werden die Boxplots der Features Area, Eccentricity, Majoraxis und der Solidity der 14 Gattungen erstellt. In der Abbildung 17 ist der Boxplot der Fläche ersichtlich. Die anderen Boxplots sind im elektronischen Anhang zu finden.

Die Kreise entsprechen, laut Definition des Boxplots, den Ausreißern. Es ist zu sehen, dass die Grafik ziemlich viele davon enthält. Es handelt sich aber auch um grosse Datensätze, wodurch der prozentuale Anteil an Ausreißern trotzdem wieder sehr klein sein könnte. Die Ausreisser deuten auf «schlechte» Daten im Datenset hin. Es ist zu erkennen, dass die Anzahl an Ausreißern abhängig von der Gattung ist und dass es einige Gattungen gibt, die scheinbar kaum Ausreisser aufweisen.

Mit schlechten Daten sind die Bilder gemeint. Die extrahierten Features der schlechten Bilder verfälschen das Messergebnis. Zudem können schlechte Bilder ein fehlerhaftes Verhalten der Feature-Extraktion verursachen. Eher unwahrscheinlich ist, dass die Biologie diese enormen Variationen verursacht und somit Grund für diese Ausreisser ist. Mit der darauffolgenden Untersuchung (Einblick in die Bilder und deren Daten) gilt es, dies herauszufinden.

Anhand der unterschiedlichen Lagen der Boxen ist die Variation zwischen den Gattungen ersichtlich. Es ist zu erkennen, dass einige Gattungen wie Fagus oder Cupressus nur schon anhand dieses Features einigermaßen gut von den anderen zu unterscheiden sind. Bei vielen Gattungen ist dieses jedoch sehr ähnlich und überschneidend mit anderen Gattungen. Eine Unterscheidung mit nur diesem Feature ist somit eher schwierig, und es braucht somit dazu noch andere Features. Bei den Features, welche eine Information über die Grösse der Polle beinhalten, also Area, Perimeter sowie Major- und Minor-Axis, wurde die grösste Variation zwischen den Arten festgestellt. Sie werden somit vermutlich auch die wichtigsten Features für die Klassifizierung der Gattungen sein.

Anhand des Boxplots ist bereits grob die Variation innerhalb einer Art ersichtlich. Es ist zu erkennen, dass diese je nach Gattung sehr unterschiedlich ist. Für eine genauere Analyse der Variation innerhalb einer Gattung eignet sich ein Histogramm besser. Dieses gibt einen genaueren Einblick in die Verteilung.

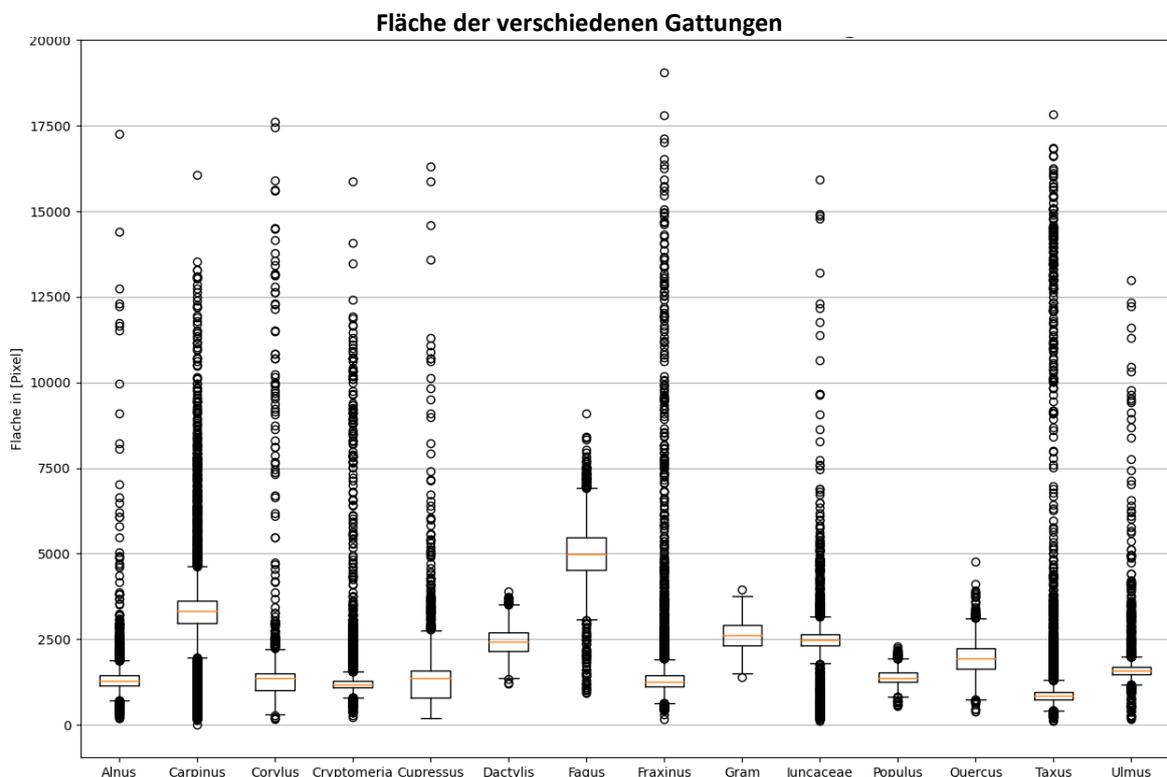


Abbildung 17: Boxplot der Fläche

3.1.3 Einblick in die Bilder und dazugehörigen Features

Der Einblick in die Bilder bestätigt die erste Annahme, dass womöglich «schlechte» Daten im Datenset vorhanden sind. Die Abbildung 18 zeigt folgende «schlechte» Bilder, welche in den Datensets gefunden wurden.

Dazu zählen: schlechter Fokus, fehlendes Teilchen, durch den Rand abgeschnittenes Teilchen, irreguläres Teilchen, zwei Teilchen auseinander und zwei aneinander. Die Auswirkungen konnten anhand der extrahierten Features des jeweiligen Bildes ausfindig gemacht werden. Dabei ist festzuhalten, dass ein stark verschwommenes Bild oder ein fehlendes Teilchen oft zu einem fehlerhaften Verhalten bei der Bildverarbeitung führt. Es wird oft eine zu grosse Fläche berechnet.

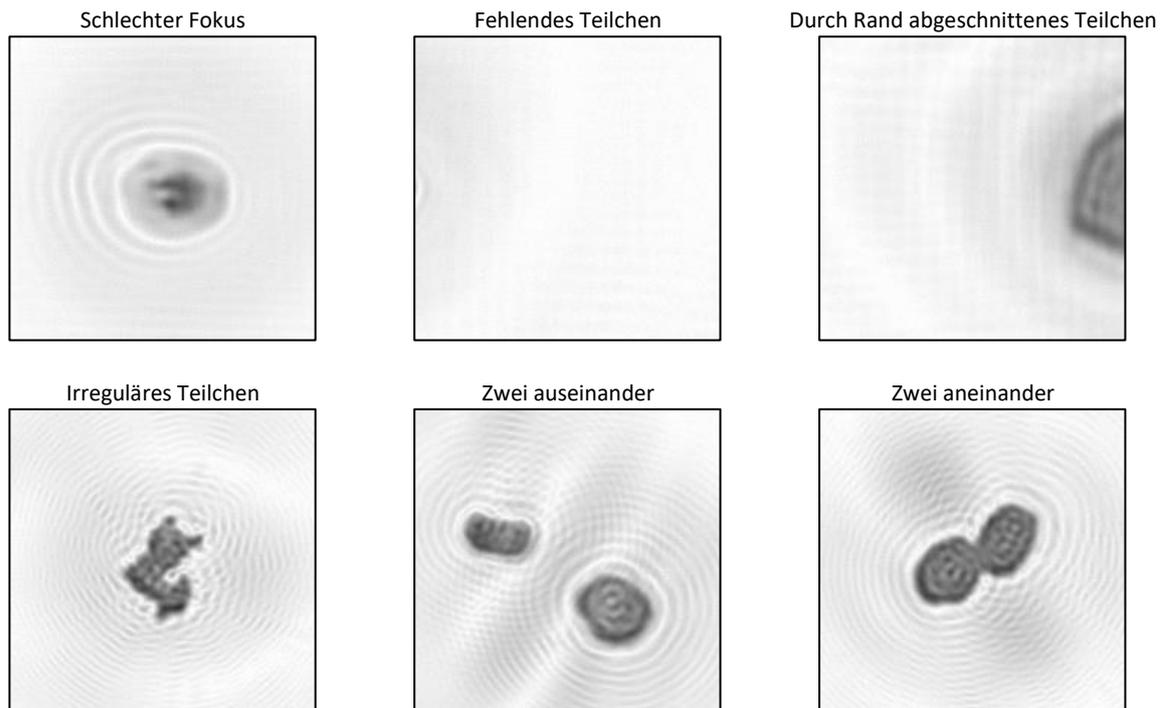


Abbildung 18: Klassen von schlechten Bildern

3.1.4 Fazit

Mit den Boxplots konnte ein erster Überblick über die Daten verschafft werden. Die Boxplots deuteten auf fehlerbehaftete Daten. Zusätzlich gaben die Boxplots bereits einen ersten Einblick in die Variation zwischen den Arten und grob in die Variation innerhalb der Arten. Somit konnte bereits eine erste Einschätzung für die Klassifizierung der Gattungen vorgenommen werden. Der anschliessende Einblick in die Bilder bestätigt die Vermutung von schlechten Bildern in den Datensets. Mögliche Fehlerquellen und ihre Auswirkungen konnten ausfindig gemacht werden. Das Ziel der Untersuchung wurde somit erfüllt. Mit dem Wissen können nun Trainings- und Testdatensets erstellt werden, welche nachfolgend kurz vorgestellt werden.

3.1.5 Erstellung Trainings- und Testdatenset

Für eine spätere Klassifikation wurde ein Trainings- sowie ein Testdatenset erstellt. Die Datensätze beinhalten rund 300 Pollendaten (Bildpaar und JSON-File) pro Gattung. Sie erhalten jeweils zwei Labels, zum einen die Gattung und zum andern die Qualität. Die Gattung wird für die spätere Klassifizierung der Gattung erforderlich sein und die Qualität für die spätere automatische Aussortierung. Die Qualität wurde, wie in der Tabelle 6 (siehe Kap. 2.3.1), gelabelt. Es wurde dabei aus eigenem Ermessen entschieden. Bei den meisten schlechten Bildern war das nicht schwierig, aber vor allem bei den verschwommenen Bildern war es schwierig abzuschätzen, was noch zu einer guten oder schlechten Qualität zählen soll. Die weiteren Untersuchungen erfolgten mit dem Trainingsdatenset und den daraus mit guter Qualität gelabelten Daten.

3.2 Genauere Analyse der Daten

Ziel dieser Untersuchung ist es, einen tieferen Einblick in die Daten des Trainingsdatensatzes zu erhalten. Es wird dabei nur mit den gut gelabelten Daten gearbeitet. Es soll mehr über die Variation innerhalb einer Art gefunden werden. Zudem sollen Abhängigkeiten von Features ausfindig gemacht werden. Wenn möglich soll bereits ein Konzept für die automatische 3D-Form-Schätzung aufgrund der Untersuchung herausgefunden werden. Es soll untersucht werden, ob allenfalls eine andere Definition von Image 0 und 1 günstiger wäre. Die Orientierung als neues Feature soll begutachtet werden.

Nach der Untersuchung soll definiert werden, welche Features für die Klassifikation der Gattungen und für die automatische Aussortierung verwendet werden sollen.

3.2.1 Vorgehen

Wie bereits erwähnt, erfolgen die folgenden Untersuchungen mit dem Trainingsdatensatz und den daraus gut gelabelten Daten.

Schritt 1: Kontrolle Bildverarbeitung und erste 3D-Form-Schätzung

In einem ersten Teil wird die korrekte Funktionsweise der Bildverarbeitung an den Trainingsdaten geprüft. Erst nach geprüfter und korrekter Funktionsweise sollen weitere Untersuchungen vorgenommen werden. Mit dieser Untersuchung kann direkt versucht werden, eine 3D-Form der Pollen abzuschätzen. Die Erkenntnisse sollen für die automatische 3D-Form-Schätzung nützlich sein. Zudem kann ein erster Vergleich mit der Referenzliste vorgenommen werden.

Schritt 2: Neue Definition von Image 0 und Image 1

Dieser Abschnitt untersucht, ob sich eine neue Definition von Image 0 und Image 1 anhand der Exzentrizität eignen könnte. Die bisherige Definition von Image 0 und Image 1 besteht anhand des Messaufbaus.

Schritt 3: Variation innerhalb der Gattung und Abhängigkeiten von Features

Dieser Teil soll einen genaueren Einblick in die Verteilung der Features und somit in die Variation innerhalb der Art geben. Zudem sollen die Abhängigkeiten zwischen den Features, speziell zwischen den Features von Image 0 und Image 1, herausgefunden werden.

Schritt 4: Untersuchung der Orientierung

In einem letzten Punkt soll das neu extrahierte Feature, die Orientierung, untersucht werden. Es soll entschieden werden, ob es in die Feature-Liste für die Klassifizierung der Gattung genommen werden soll oder nicht.

3.2.2 Kontrolle Bildverarbeitung und erste 3D-Form-Schätzung

Damit kontrolliert werden kann, ob die Bildverarbeitung bei den Daten aus dem Trainingsset auch erwartungsgemäss funktioniert, wird bei allen Bildpaaren jeweils die gefittete Ellipse eingezeichnet. Ein Beispiel ist in der Abbildung 19 ersichtlich. Anhand der gefitteten Ellipse kann bereits gut eingeschätzt werden, ob die Feature-Extraktion richtig funktioniert. Das Ganze wird für alle Bildpaare aus dem Trainingsset gemacht und einzeln kontrolliert.

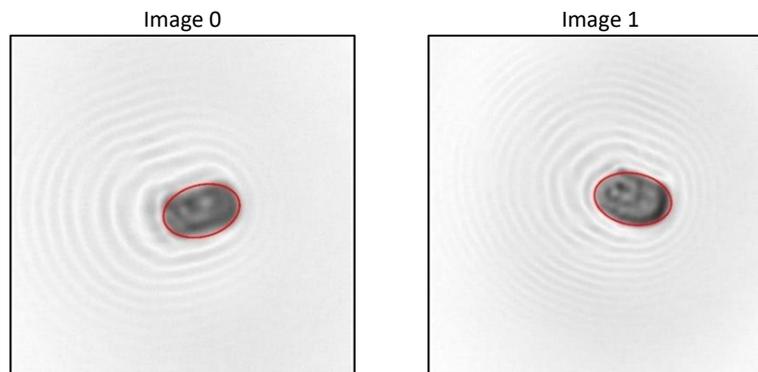


Abbildung 19: Ellipsen-Fit

Die Feature-Extraktion hat bei allen Bildern aus dem Trainingsatz wunschgemäss gearbeitet. Nun kann mit der weiteren Untersuchung fortgefahren werden.

Anhand der beiden gefitteten Ellipsen konnte bereits eine erste 3D-Form-Schätzung vorgenommen werden. Dabei wurde versucht, mit den beiden Ellipsen ein Rotationsellipsoid abzuschätzen sowie eine für eine Gattung meistauftretende Form zu finden. Die Ergebnisse sind in der Tabelle 11 ersichtlich. Zusätzlich wurden jeweils noch die Major- und Minor-Axis-Mediane und die jeweiligen Breiten der Boxen (25 – 75 Quantile) miteingetragen. Mit dieser Tabelle konnte ein erster Vergleich mit der Referenzliteratur vorgenommen werden.

	Geschätzte 3D-Form aus den Bildern	Major-Axis [um]	Breite Box [um]	Minor-Axis [um]	Breite Box [um]
Alnus	oblate	26.46	2.7	16.74	2.7
Carpinus	isodiametric, prolate	37.26	2.7	32.94	3.24
Corylus	isodiametric, oblate	24.3	2.16	20.52	2.16
Cryptomeria	oblate	25.92	2.16	16.2	2.16
Cupressus	prolate (und die andern beiden)	23.22	5.94	20.52	8.1
Dactylis	prolate (Dreiecksform teils)	35.64	4.86	25.92	4.32
Fagus	prolate	47.52	7.02	37.8	4.32
Fraxinus	prolate	25.38	4.05	17.28	2.7
Gram	prolate (oblate)	35.1	4.32	27	4.32
Juncaceae	oblate, prolate	32.4	3.24	27.54	2.16
Populus	oblate, prolate	25.92	3.24	19.44	3.24
Quercus	prolate	35.64	8.64	19.44	3.24
Taxus	isodiametric (und die andern beiden)	18.9	2.7	10.26	2.7
Ulmus	oblate	17.28	2.16	19.98	1.62

Tabelle 11: 3D-Form-Schätzung

Wie die Tabelle zeigt, war nicht immer eine klare 3D-Form von einer Gattung auszumachen. Zudem ist zu sagen, dass die 3D-Form-Schätzung anhand der beiden Ellipsen nicht immer ganz so einfach war. Dies zeigt, dass die automatische 3D-Form-Schätzung wahrscheinlich eine nicht ganz so einfache Aufgabe sein wird. Es konnte leider noch kein Konzept dafür gefunden werden. Die Grössen stimmen bis auf Ulmus gut mit der Referenzliste überein. Bei der Form gibt es zum Teil Abweichungen.

3.2.3 Neue Definition von Image 0 und Image 1

Die bisherige Definition von Image 0 und 1 war anhand des Messaufbaus bestimmt. Folgende Untersuchung soll zeigen, ob eine neue Definition anhand der Exzentrizität geeigneter ist oder nicht. Es wird jeweils dem Image, welches eine kleinere Exzentrizität aufweist, also rundlicher ist, die Nummer 0 zugewiesen und dem anderen die Nummer 1. Dies ist jeweils am Feature mit der Beschriftung `_0` und `_1` ersichtlich.

Die Abbildung 20 zeigt die Histogramme der Fläche von *Alnus*, *Carpinus* und *Quercus* von Image 0 und Image 1. Links ist das Feature anhand des Messaufbaus definiert und rechts neu anhand der Exzentrizität.

Die drei Gattungen wurden aufgrund ihrer Form ausgewählt. Bei *Alnus* handelt es sich um eine Gattung, welche meist eine oblate Form aufweist. *Carpinus* ist meist isodiametrisch, und *Quercus* weist eine stark prolate Form auf.

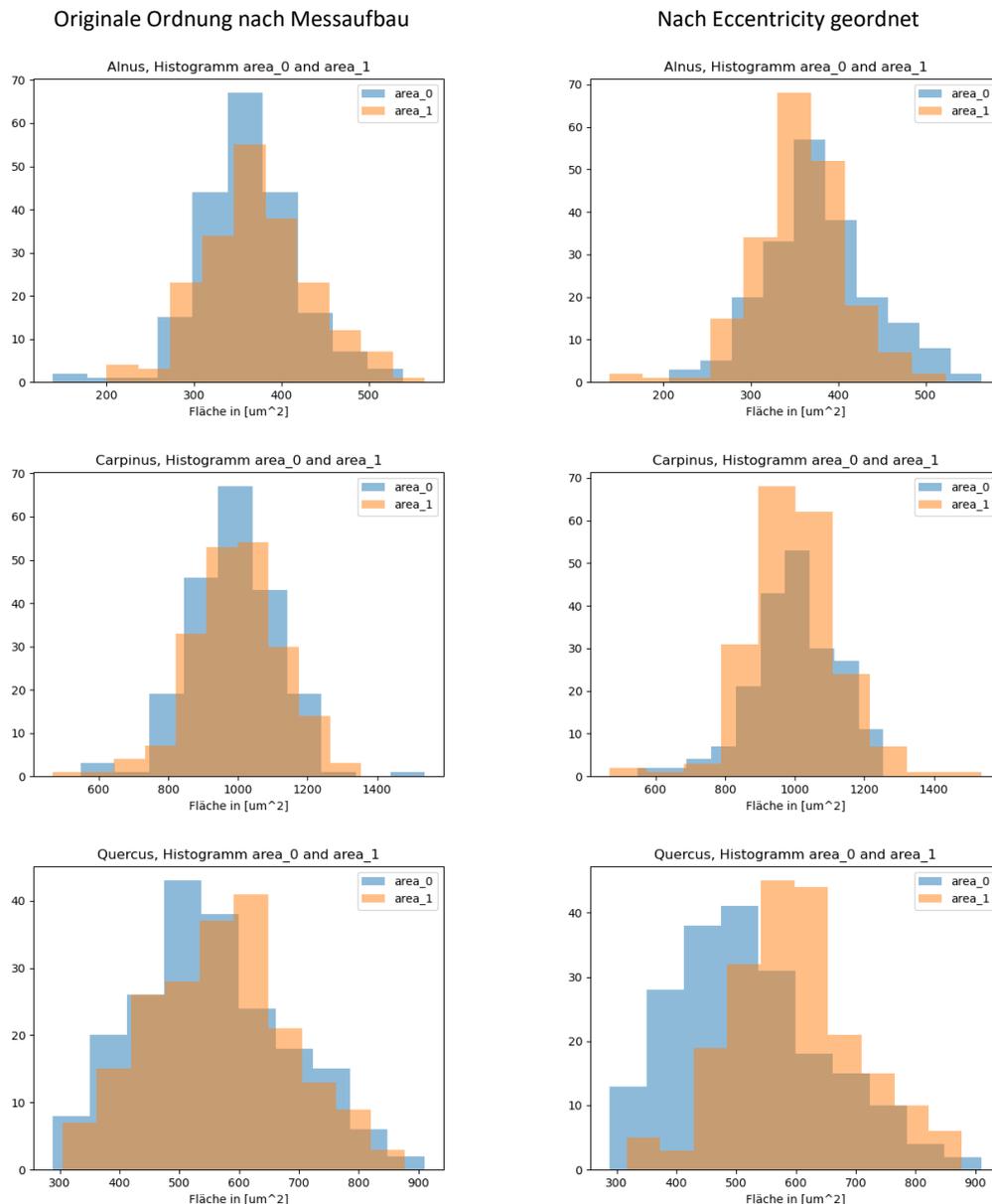


Abbildung 20: Definition von Image 0 und Image 1

Bei den linken Histogrammen ist zu sehen, dass die Verteilung der Features von Image 0 und Image 1 anhand der bisherigen Definition ziemlich gleich aussieht.

Bei den rechten Histogrammen von *Alnus* und *Quercus* ist zu erkennen, dass die beiden Verteilungen sich ein wenig voneinander verschieben. Bei *Alnus* verschiebt sich die blaue Kurve nur ganz leicht nach rechts und bei *Quercus* gut ersichtlich nach links. Die Verschiebung lässt sich aufgrund der Form der Polle erklären. Das Beispiel mit *Quercus*: *Quercus* ist stark prolate. Die rundere Form auf dem Bild ist somit kleiner als die

stärker elliptische Form. Bei Alnus ist gerade das Umgekehrte der Fall. Bei Carpinus ist keine Verschiebung ersichtlich, und die Kurven liegen ziemlich gleich aufeinander. Das liegt daran, dass Carpinus meist eine isodiametrische Form aufweist. Für eine 3D-Form-Bestimmung genügt diese Erkenntnis wohl kaum, aber es ist zu sehen, dass die Kurven nach der neuen Definition dadurch teils ein wenig schmaler werden können. Dies bedeutet eine kleinere Variation innerhalb einer Gattung, was für die Klassifikation wünschenswerter ist.

Deswegen werden neu Image 0 und 1 nicht mehr nach dem Messaufbau definiert, sondern nach der Exzentrizität.

3.2.4 Variation innerhalb Art, Abhängigkeit, Mittelwerte

Am Boxplot ist die Variation innerhalb einer Gattung nur grob ersichtlich. Um einen besseren Einblick in die Verteilung zu erhalten, ist ein Histogramm hilfreicher.

Es wurden von jeder Gattung einzelne Histogramme von allen Features erstellt. Ein Beispiel ist in der Abbildung 21 ersichtlich. Hierbei handelt es sich um die Major-Axis von Image 1 und 2.

Die Features aus Image 0 und 1 wurden jeweils zusammengenommen. Zusätzlich wurde noch ein Scatterplot (siehe Abb. 21 rechts) erstellt. Dieses dient dazu, um allenfalls einen linearen oder sonstigen Zusammenhang festzustellen. (Die Grafiken von allen Gattungen und Features sind im elektronischen Anhang ersichtlich.)

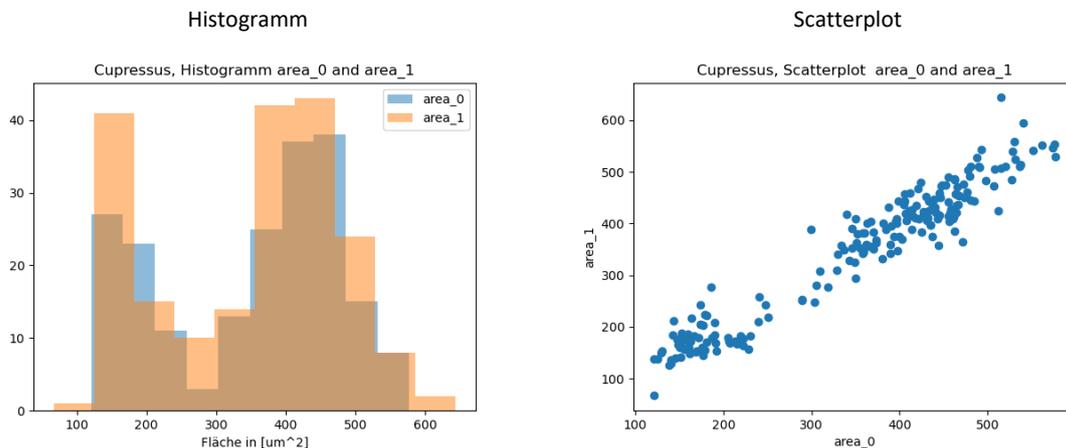


Abbildung 21: Histogramm und Scatterplot

Als Erstes ist zu bemerken, dass es zu viele Grafiken für eine genaue und einzelne Analyse sind. Es sind jeweils 126 (14*9) Histogramme und Scatterplots. Deswegen wurde nur oberflächlich nach Besonderheiten Ausschau gehalten. Bei Corylus und Cupressus wurden, wie in der Abbildung ersichtlich, jeweils zwei Berge bei der Verteilung ausfindig gemacht. Das liegt darin, dass sich zwei unterschiedliche Größen von Pollen im Datenset befinden. Anhand des Scatterplots war jeweils nur grob eine Korrelation zwischen den Features auszumachen. Für eine genauere Analyse wurden Korrelationsmatrizen erstellt. An diesen konnte nun genau festgestellt werden, ob eine Korrelation zwischen den Features vorhanden ist. Die Korrelationsmatrizen sind im elektronischen Anhang zu finden.

Wie erwartet, wurde eine Korrelation zwischen den Features auf Image 0 und 1 festgestellt. Zudem wurden auch starke Korrelationen zwischen den Features, welche Informationen über die Grösse beinhalten, festgestellt, wie zum Beispiel zwischen Area und Perimeter. Dies zeigt eine gewisse Redundanz zwischen den Features.

Für ein späteres Standardisieren des Datensets wurden noch die Mittelwerte und Standardabweichungen der einzelnen Features über alle Gattungen im Trainingsset ausfindig gemacht. Die Ergebnisse sind in der Tabelle 12 ersichtlich. Diese Daten werden immer, wenn es um das Standardisieren geht, verwendet.

Feature	Mittelwert μ	Standardabweichung σ	Einheiten
solidity_0	0.966	0.015	Zahl von 0 - 1
solidity_1	0.963	0.018	Zahl von 0 - 1
area_0	573.528	323.872	μm^2
area_1	584.292	336.307	μm^2
minorAxis_0	23.618	7.008	μm
minorAxis_1	22.572	6.986	μm
majorAxis_0	29.382	7.438	μm
majorAxis_1	31.365	8.134	μm
perimeter_0	92.383	27.404	μm
perimeter_1	94.259	28.157	μm
maxIntensity_0	0.709	0.010	Zahl von 0 - 1
maxIntensity_1	0.708	0.010	Zahl von 0 - 1
minIntensity_0	0.251	0.053	Zahl von 0 - 1
minIntensity_1	0.245	0.056	Zahl von 0 - 1
meanIntensity_0	0.498	0.019	Zahl von 0 - 1
meanIntensity_1	0.497	0.020	Zahl von 0 - 1
eccentricity_0	0.563	0.159	Zahl von 0 - 1
eccentricity_1	0.670	0.131	Zahl von 0 - 1

Tabelle 12: Mittelwerte und Standardabweichungen der Features

3.2.5 Untersuchung der Orientierung

Das neu extrahierte Feature soll untersucht werden. Es soll entschieden werden, ob es in die Feature-Liste für die Klassifizierung der Gattungen miteinbezogen werden soll oder nicht. In der Abbildung 22 ist der Boxplot der Orientierung der 14 Gattungen ersichtlich. Die Definition der Orientierung ist im Theorieteil 2.2.2 zu finden.

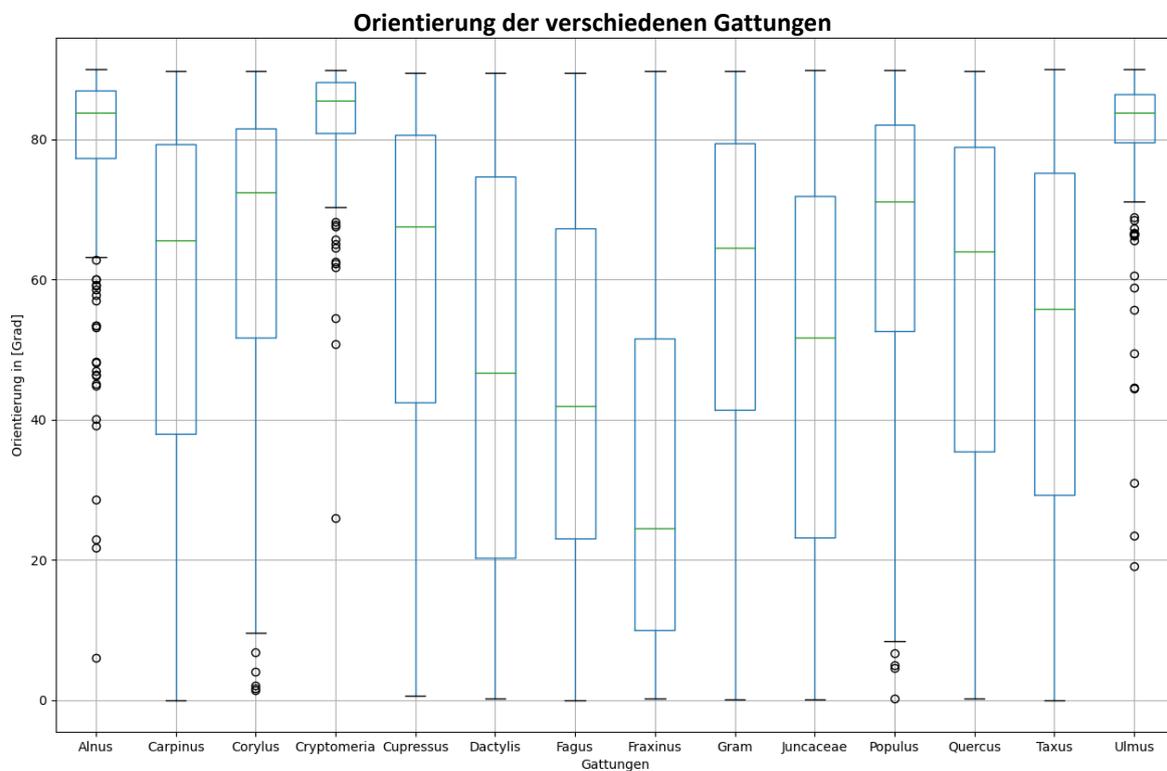


Abbildung 22: Boxplot-Orientierung

Bei den meisten Gattungen ist die Orientierung über die 90° gleich verteilt. Es ist somit kein spezielles Flugverhalten zu erkennen. Bei den drei Gattungen Alnus, Cryptomeria und Ulmus jedoch neigt das Flugverhalten dazu nahe an die 90° . Dies bedeutet, dass «die Ellipse» horizontal in der Luft liegt. Bei diesen drei Gattungen handelt es sich um eine oblate Polle. Es ist somit wahrscheinlich, dass das Flugverhalten aufgrund der Form verursacht wird.

Da jedoch bei den meisten Gattungen das Flugverhalten gleich verteilt ist, wird dieses Feature nicht in die Feature-Liste für die Klassifizierung der Gattungen miteinbezogen. Zu einem späteren Zeitpunkt könnte es vielleicht trotzdem in die Liste aufgenommen werden, um zu untersuchen, ob es für die Klassifikation einen Nutzen hat.

3.2.6 Fazit

In einem ersten Teil konnte die bisherige Feature-Extraktion auf korrektes Verhalten bei den «guten» Bildern im Trainingsdatensatz geprüft werden. Zusätzlich konnte mit dem Ellipsenfit eine erste grobe Schätzung eines Rotationsellipsoids gemacht werden. Diese 3D-Form-Schätzung hat sich aber als nicht ganz so einfach herausgestellt. Es konnte noch kein «Konzept» für eine automatische 3D-Form-Schätzung entwickelt werden.

Mit der neuen Definition von Image 0 und Image 1 konnte die Variation innerhalb der Art leicht verkleinert werden. Zusätzlich ist aus den Verteilungen von der Fläche schon leicht die Form zu erkennen. Es ist somit eine mögliche geeignetere Definition von Image 0 und 1 gefunden worden als die bisherige. Dies sollte jedoch im Hinterkopf behalten werden, da es womöglich auf andere Dinge noch unbekannte Nebeneffekte haben könnte.

Eine genauere Betrachtung der Verteilung der Features und deren Abhängigkeiten gaben einen tieferen Einblick in die Daten. Es hat sich aber auch gezeigt, dass eine solche Untersuchung schnell sehr aufwendig und gross werden und somit der Überblick verloren gehen kann. Zudem wurde eine Tabelle der Mittelwerte und Standardabweichungen erstellt, welche für späteres Standardisieren der Daten gebraucht werden kann. In einem letzten Punkt konnte noch das neu extrahierte Feature, die Orientierung, untersucht werden. Es ist dabei bei drei Gattungen ein besonderes Flugverhalten festgestellt worden. Da jedoch bei den anderen Gattungen kein spezielles Flugverhalten ermittelt wurde, werden diese Features vorerst nicht in die Liste der Features für die Klassifizierung der Gattungen miteinbezogen.

Bis auf das Finden eines Konzepts für die 3D-Form-Schätzung konnten die Ziele der Untersuchung erfüllt werden.

Die Features, welche für die Klassifikation der Gattungen verwendet werden, sind nun definiert. Es handelt sich um die Features, welche bereits in der bisherigen Feature-Extraktion extrahiert wurden. Diese werden neu nach der Exzentrizität geordnet. Die neu hinzugefügten Features wie die Maximalwerte und Schwerpunkte werden nicht in die Liste aufgenommen, da diese keine neue Information enthalten, welche für die Klassifizierung der Gattungen nützlich ist.

3.3 Feature-Reduktion mittels PCA

Das bisherige Datenset umfasst 18 Features. Vorherige Untersuchungen haben gezeigt, dass Redundanz zwischen den Features steckt. Es sollte somit möglich sein, die Daten-Dimensionalität, also die Anzahl Features, zu reduzieren, ohne einen merklichen Informationsverlust einzubüßen. PCA ist ein gutes Werkzeug dafür. Zusätzlich können mit PCA Visualisierungen erstellt werden, welche bereits erste Abschätzungen für die Unterscheidung der Arten zulassen.

Die Tabelle 13 zeigt nochmals die Struktur des Datensets mit den 18 originalen Features auf, welche für die Klassifizierung verwendet werden sollen. Bei den Zeilen handelt es sich, wie bereits im Theorieteil vorgestellt, um die Anzahl Samples. Die Einträge der Tabelle entsprechen der Datenmatrix, wobei m1 der Anzahl Samples und n1 der Anzahl Features entspricht. Als Hilfestellung ist die Tabelle 14 für die Notation, welche im Theorieteil (siehe Kap. 2.4) bereits vorgestellt wurde, noch einmal aufgeführt.

m1 = Anzahl Samples	Index/Sample	solidity_0	solidity_1	area_0	area_1	minorAxis_0	minorAxis_1	majorAxis_0	majorAxis_1	perimeter_0	perimeter_1	maxIntensity_0	maxIntensity_1	minIntensity_0	minIntensity_1	meanIntensity_0	meanIntensity_1	eccentricity_0	eccentricity_1	
	1																			
	2																			
	3																			
	⋮																			
	n1 = Anzahl Features																			

Tabelle 13: Struktur Datensatz mit 18 Features

X	Datenmatrix	m1 = Anzahl Samples n1 = Anzahl originale Features
P	Transformationsmatrix	m2 = Anzahl PCA Features n2 = Anzahl originale Features
Y	Transformierte Datenmatrix	m3 = Anzahl Samples n3 = Anzahl PCA Features

Tabelle 14: Notation PCA

3.3.1 Vorgehen

Gearbeitet wird mit gut gelabelten Daten aus dem Trainingsdatenset. Das Vorgehen ist wie folgt aufgebaut:

Schritt 1: Kovarianzmatrix von X

In einem ersten Schritt wird die Kovarianzmatrix der standardisierten Datenmatrix berechnet.

Schritt 2: Transformationsmatrix P

Mit der Kovarianzmatrix kann nun die Transformationsmatrix P erstellt werden.

Schritt 3: Zeilenreduktion der Transformationsmatrix P

Anhand der Eigenwerte kann eine Zeilenreduktion von P vorgenommen werden. Mit der reduzierten Transformationsmatrix lässt sich nun die Dimension des Datensatzes reduzieren.

Schritt 4: Kovarianzmatrix von Y

Die Kovarianzmatrix von Y soll nun die Effekte von PCA verdeutlichen.

Schritt 5: Visualisierung mit PCA

In einem letzten Schritt wird noch eine Visualisierung mit den ersten beiden PCA Features vorgenommen.

3.3.2 Kovarianzmatrix von X

Es wird die Kovarianzmatrix der standardisierten Datenmatrix berechnet. Die Rechnung der Standardisierung sieht, wie bereits im Theorieteil vorgestellt, wie folgt aus. Es werden dabei die Mittelwerte und Standardabweichungen aus der Tabelle 12 (siehe Kap. 3.2.4) verwendet.

$$X = \frac{X_{unscaled} - \mu}{\sigma}$$

Anschliessend wird die Kovarianzmatrix von X berechnet. Diese ist in der Abbildung 23 ersichtlich. Als Hilfe wurden die jeweiligen 18 originalen Features hinzugefügt. Für eine bessere Übersicht wurden die Werte auf eine signifikante Stelle nach dem Komma gerundet. Dies gilt für alle folgenden Matrizen, die in diesem Kapitel erläutert werden.

Da X normal standardisiert ist, handelt es sich bei der Kovarianzmatrix zugleich um die Korrelationsmatrix. Die diagonalen Werte entsprechen der Varianz der Features, und die nicht diagonalen Werte entsprechen der Kovarianz der Features.

Anhand der Kovarianzen ist ersichtlich, dass Redundanz zwischen den Features steckt. Da diese meistens ungleich null sind. Je grösser die Kovarianz, desto stärker korrelieren die Features. Es ist gut ersichtlich, dass vor allem die gleichen Features von Image 0 und 1 stark miteinander korrelieren. Die Abhängigkeit von Image 0 und 1 war aufgrund der nicht zu komplizierten Form einer Polle zu erwarten. Zudem gut ersichtlich ist, dass die Features, welche etwas über die Grösse der Polle aussagen, also Area, Perimeter, Major- und Minor-Axis, stark miteinander korrelieren.

Da viel Redundanz in den Daten vorhanden ist, ist davon auszugehen, dass mit PCA nur schon durch das Eliminieren von Redundanz die Dimension um einiges reduziert werden kann.

solidity		area		minor-ax		major-ax		perimeter		max-Int.		min-Int.		mean-Int.		ecc.	
im0	im1	im0	im1	im0	im1	im0	im1	im0	im1	im0	im1	im0	im1	im0	im1	im0	im1
1.0	0.2	0.2	0.2	0.2	0.2	0.1	0.1	0.0	0.1	-0.3	-0.1	0.2	0.1	-0.1	0.0	-0.1	-0.1
0.2	1.0	0.3	0.3	0.3	0.3	0.2	0.2	0.2	0.1	-0.1	-0.3	0.1	0.2	0.1	-0.1	-0.2	-0.3
0.2	0.3	1.0	0.9	1.0	0.9	1.0	0.9	0.9	0.9	-0.1	-0.2	0.1	0.1	0.4	0.3	-0.2	-0.3
0.2	0.3	0.9	1.0	0.9	1.0	0.9	0.9	0.9	0.9	-0.2	-0.1	0.1	0.1	0.3	0.3	-0.2	-0.3
0.2	0.3	1.0	0.9	1.0	1.0	0.9	0.8	0.9	0.9	-0.1	-0.1	0.1	0.1	0.4	0.3	-0.4	-0.4
0.2	0.3	0.9	1.0	1.0	1.0	0.9	0.8	0.9	0.9	-0.1	-0.1	0.1	0.2	0.3	0.4	-0.4	-0.5
0.1	0.2	1.0	0.9	0.9	0.9	1.0	0.9	0.9	0.9	-0.1	-0.2	0.1	0.1	0.4	0.2	0.0	-0.1
0.1	0.2	0.9	0.9	0.8	0.8	0.9	1.0	0.8	0.9	-0.2	-0.2	0.0	0.1	0.3	0.3	0.0	0.0
0.0	0.2	0.9	0.9	0.9	0.9	0.9	0.8	1.0	0.9	0.0	-0.1	0.0	0.1	0.4	0.3	-0.2	-0.2
0.1	0.1	0.9	0.9	0.9	0.9	0.9	0.9	0.9	1.0	-0.1	-0.1	0.0	0.1	0.3	0.4	-0.2	-0.2
-0.3	-0.1	-0.1	-0.2	-0.1	-0.1	-0.1	-0.2	0.0	-0.1	1.0	0.2	0.2	0.0	0.9	0.1	-0.1	0.0
-0.1	-0.3	-0.2	-0.1	-0.1	-0.1	-0.2	-0.2	-0.1	-0.1	0.2	1.0	0.0	0.2	0.1	0.9	0.0	0.0
0.2	0.1	0.1	0.1	0.1	0.1	0.1	0.0	0.0	0.0	0.2	0.0	1.0	0.1	0.2	0.0	-0.1	-0.1
0.1	0.2	0.1	0.1	0.1	0.2	0.1	0.1	0.1	0.1	0.0	0.2	0.1	1.0	0.0	0.3	-0.1	-0.2
-0.1	0.1	0.4	0.3	0.4	0.3	0.4	0.3	0.4	0.3	0.9	0.1	0.2	0.0	1.0	0.2	-0.2	-0.2
0.0	-0.1	0.3	0.3	0.3	0.4	0.2	0.3	0.3	0.4	0.1	0.9	0.0	0.3	0.2	1.0	-0.2	-0.2
-0.1	-0.2	-0.2	-0.2	-0.4	-0.4	0.0	0.0	-0.2	-0.2	-0.1	0.0	-0.1	-0.1	-0.2	-0.2	1.0	0.8
-0.1	-0.3	-0.3	-0.3	-0.4	-0.5	-0.1	0.0	-0.2	-0.2	0.0	0.0	-0.1	-0.2	-0.2	-0.2	0.8	1.0

Abbildung 23: Kovarianzmatrix von X mit Features als Hilfe

3.3.3 Transformationsmatrix P

Es können nun die Eigenwerte und Eigenvektoren der Kovarianzmatrix berechnet werden. Mit den anhand der Eigenwerte geordneten Eigenvektoren kann nun die Transformationsmatrix P erstellt werden. Diese ist in der Abbildung 24 ersichtlich. Die Zeilen entsprechen somit den Eigenvektoren aus der Kovarianzmatrix von X. Die Werte entsprechen der Gewichtung des originalen Features für das neue PCA Feature.

Die Eigenwerte sind der Größe nach geordnet. Sie gehören zu den Eigenvektoren der Zeilen 1 bis 18 der Transformationsmatrix P. Sie entsprechen der Varianz der neuen PCA Features. Die Eigenwerte sind in der Abbildung 24 ersichtlich. Zudem wird die Varianz noch in Prozenten angegeben. Es ist ersichtlich, dass die letzten acht PCA Features fast keine Information mehr enthalten und somit ohne grossen Informationsverlust eliminiert werden können.

Transformationsmatrix P																		
0.1	0.1	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.0	0.0	0.0	0.0	0.1	0.1	-0.1	-0.1
-0.2	-0.1	-0.1	-0.1	0.0	0.0	-0.1	-0.1	0.0	0.0	0.4	0.5	0.1	0.2	0.4	0.4	-0.2	-0.2	
-0.3	-0.4	0.0	0.0	-0.1	-0.1	0.2	0.2	0.1	0.1	0.1	0.2	-0.2	-0.1	0.1	0.2	0.5	0.5	
-0.2	0.1	0.0	0.0	0.0	0.0	0.1	0.0	0.1	-0.1	0.5	-0.4	0.2	-0.3	0.5	-0.4	0.0	0.1	
-0.5	-0.1	0.0	0.0	0.1	0.1	-0.1	-0.1	0.1	0.0	-0.1	-0.1	-0.6	-0.3	-0.1	-0.1	-0.3	-0.3	
0.4	-0.5	0.0	0.0	0.1	0.0	0.0	0.0	-0.1	0.1	-0.1	0.1	0.3	-0.7	0.0	0.1	-0.2	-0.1	
-0.6	-0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	-0.2	-0.1	0.6	0.2	-0.2	-0.1	-0.1	-0.1	
0.3	-0.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	-0.2	-0.3	0.5	0.1	-0.2	-0.1	0.0	
-0.1	0.1	-0.2	0.2	-0.1	0.0	-0.3	0.4	-0.3	0.3	0.1	-0.1	0.0	0.0	0.0	0.0	-0.5	0.5	
0.1	0.0	0.3	-0.2	0.3	-0.3	0.1	-0.1	0.4	-0.4	-0.1	0.1	0.0	0.0	0.0	0.0	-0.4	0.5	
0.1	0.0	-0.4	-0.4	-0.1	-0.1	-0.1	0.0	0.6	0.5	-0.1	-0.1	0.1	0.1	0.1	0.1	0.0	0.0	
0.1	0.0	0.2	0.4	-0.2	0.0	-0.4	-0.3	0.4	0.2	0.3	0.2	0.0	0.0	-0.3	-0.2	0.1	0.1	
-0.1	0.1	0.3	-0.2	0.2	-0.2	0.2	-0.4	-0.4	0.6	0.0	0.1	0.0	0.0	0.0	-0.1	0.0	0.1	
0.0	0.0	0.1	-0.1	-0.3	-0.4	0.4	0.5	0.0	0.0	0.3	0.1	0.0	0.0	-0.3	-0.1	-0.1	-0.3	
0.0	0.0	-0.1	0.2	-0.2	-0.1	0.0	0.1	0.0	0.0	-0.4	0.5	0.0	0.0	0.5	-0.5	0.0	-0.1	
0.0	0.0	0.3	0.3	-0.3	-0.4	-0.2	-0.1	0.0	0.0	-0.3	-0.4	0.0	0.0	0.3	0.4	0.0	-0.1	
0.0	0.0	-0.3	0.3	-0.5	0.2	0.6	-0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-0.2	0.2	
0.0	0.0	-0.5	0.5	0.5	-0.5	0.1	-0.1	0.0	0.0	0.0	0.0	0.0	0.0	-0.1	0.0	0.1	-0.1	

Eigenwerte																	
7.9	2.4	2.0	1.8	1.2	1.0	0.7	0.6	0.3	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Explained-Varianz (in %)																	
0.44	0.13	0.11	0.1	0.06	0.05	0.04	0.03	0.01	0.01	0	0	0	0	0	0	0	0

Abbildung 24: Transformationsmatrix P und Eigenwerte

Die Grafik verdeutlicht noch einmal die Varianz, abhängig von der Anzahl PCA Features. Es ist zu sehen, dass mit zehn PCA Features schon fast 100 % erreicht werden. Deswegen werden für die weitere Untersuchung zehn PCA Features verwendet.

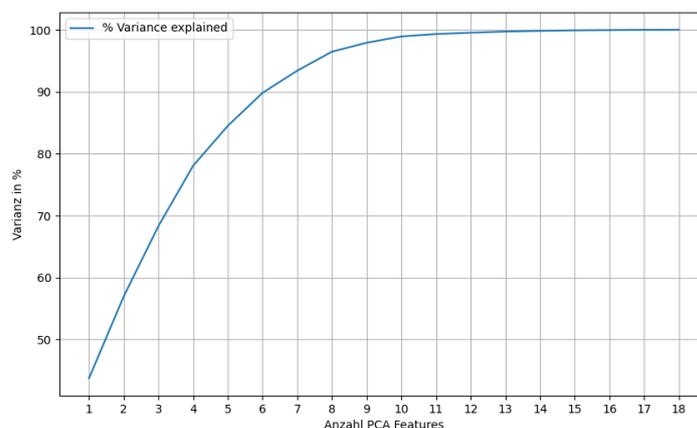


Abbildung 25: Abhängigkeit der Varianz anhand der Anzahl PCA Features

3.3.4 Zeilenreduktion der Transformationsmatrix P

Die Transformationsmatrix, welche die Feature-Zahl von 18 auf 10 reduziert, ist in der Abbildung 26 ersichtlich. Es handelt sich um die gleiche Transformationsmatrix P wie im vorherigen Kapitel. Jedoch beinhaltet sie nur noch die ersten zehn Zeilen. Die Angabe der Features dient als Hilfe.

Kovarianzmatrix mit Features als Hilfe																	
solidity		area		minor-ax		major-ax		perimeter		max-Int.		min-Int.		mean-Int.		ecc.	
im0	im1	im0	im1	im0	im1	im0	im1	im0	im1	im0	im1	im0	im1	im0	im1	im0	im1
0.1	0.1	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.0	0.0	0.0	0.0	0.1	0.1	-0.1	-0.1
-0.2	-0.1	-0.1	-0.1	0.0	0.0	-0.1	-0.1	0.0	0.0	0.4	0.5	0.1	0.2	0.4	0.4	-0.2	-0.2
-0.3	-0.4	0.0	0.0	-0.1	-0.1	0.2	0.2	0.1	0.1	0.1	0.2	-0.2	-0.1	0.1	0.2	0.5	0.5
-0.2	0.1	0.0	0.0	0.0	0.0	0.1	0.0	0.1	-0.1	0.5	-0.4	0.2	-0.3	0.5	-0.4	0.0	0.1
-0.5	-0.1	0.0	0.0	0.1	0.1	-0.1	-0.1	0.1	0.0	-0.1	-0.1	-0.6	-0.3	-0.1	-0.1	-0.3	-0.3
0.4	-0.5	0.0	0.0	0.1	0.0	0.0	0.0	-0.1	0.1	-0.1	0.1	0.3	-0.7	0.0	0.1	-0.2	-0.1
-0.6	-0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	-0.2	-0.1	0.6	0.2	-0.2	-0.1	-0.1	-0.1
0.3	-0.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	-0.2	-0.3	0.5	0.1	-0.2	-0.1	0.0
-0.1	0.1	-0.2	0.2	-0.1	0.0	-0.3	0.4	-0.3	0.3	0.1	-0.1	0.0	0.0	0.0	0.0	-0.5	0.5
0.1	0.0	0.3	-0.2	0.3	-0.3	0.1	-0.1	0.4	-0.4	-0.1	0.1	0.0	0.0	0.0	0.0	-0.4	0.5

Eigenwerte									
7.88	2.38	2.04	1.75	1.16	0.96	0.65	0.55	0.26	0.18

Explained-Varianz in %									
0.44	0.13	0.11	0.1	0.06	0.05	0.04	0.03	0.01	0.01

Abbildung 26: PCA 10

Die ersten vier PCA Features beinhalten bereits fast 70 % des Informationsgehaltes. Anhand der Werte der Eigenvektoren ist die Gewichtung der originalen Features für die neuen PCA Features ersichtlich. Die Tabelle listet die wichtigsten originalen Features für die ersten drei PCA Features auf. Dabei ist ersichtlich, dass PCA Feature 1 mit 44 % mit Abstand am meisten Informationen beinhaltet. Es ist zu erkennen, dass es sich dabei vor allem um Features, welche eine Aussage über die Grösse der Polle beinhalten, handelt. Bereits in den ersten Untersuchungen wurde festgestellt, dass darin wahrscheinlich die grösste Information steckt. Beim PCA Feature 2 sind es Max- und Mean-Intensity. Beim PCA Feature 3 ist es die Eccentricity und die Solidity. Dies sind Features, die hauptsächlich etwas über die Form der Polle beinhalten.

Feature	Explained Variance	Original Features
PCA 1	44 %	Area Minor-Axis Major-Axis Perimeter
PCA 2	13 %	Max-Intensity Mean-Intensity
PCA 3	11 %	Eccentricity Solidity

Tabelle 15: Die ersten drei PCA Features

Mit der reduzierten Transformationsmatrix kann nun die Dimensionalität des originalen Datensatzes reduziert werden. Die Rechnung verdeutlicht dies. Die Dimensionalität wird nun von 18 auf 10 reduziert.

$$\begin{matrix} m1 \\ \left[\begin{array}{c} \mathbf{X} \\ \hline \end{array} \right] \\ n1 \end{matrix} * \begin{matrix} m2 \\ \left[\begin{array}{c} \mathbf{P}^T \\ \hline \end{array} \right] \\ n2 \end{matrix} = \begin{matrix} m1 \\ \left[\begin{array}{c} \mathbf{Y} \\ \hline \end{array} \right] \\ n2 \end{matrix}$$

Die folgende Abbildung 27 symbolisiert die Transformation des originalen Datensatzes mit den 18 originalen Features auf den neuen Datensatz mit zehn PCA Features. Bei den PCA Features handelt es sich somit um Linearkombinationen der bisherigen Features. Im neuen Datensatz sind immer noch fast 100 % der Varianz enthalten.

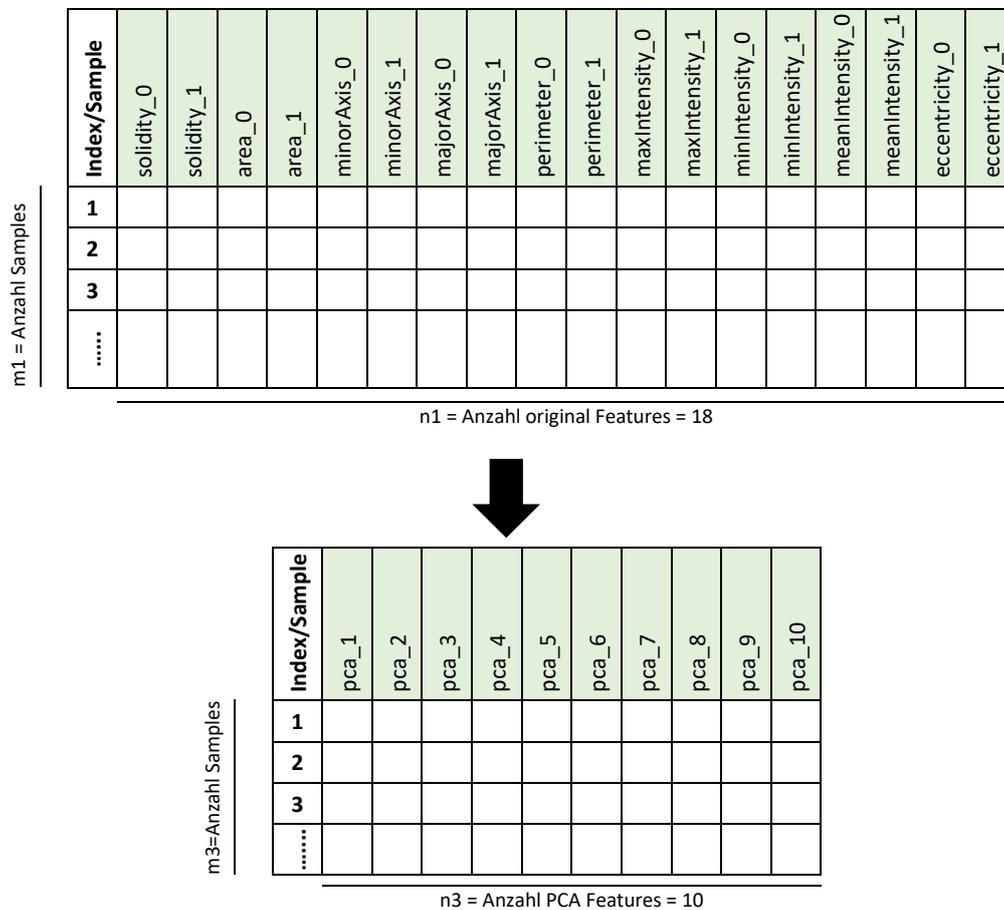


Abbildung 27: Transformation des Datensatzes

3.3.5 Kovarianzmatrix von Y

In der Abbildung ist nun die Kovarianzmatrix des reduzierten Datensatzes ersichtlich. Es handelt sich um eine 10*10 Matrix. Dies entspricht der Anzahl PCA Features. Die Diagonale beinhaltet die Varianz der Features. Dabei ist gut ersichtlich, dass diese durch PCA der Grösse nach geordnet sind. Auch gut ersichtlich ist, dass die Kovarianzen gleich null sind. Dies zeigt, dass somit keine Redundanz zwischen den PCA Features vorhanden ist.

7.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	2.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	1.8	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	1.2	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.6	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.6	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2

Abbildung 28: Kovarianzmatrix von Y

3.3.6 Visualisierung mit PCA

Die unten abgebildete Grafik wurde mit den ersten beiden PCA Features erstellt. Diese beiden Features beinhalten ca. 60 % der Varianz. Die Grafik soll bereits ein erstes Gefühl für die Unterscheidung der Gattungen geben.

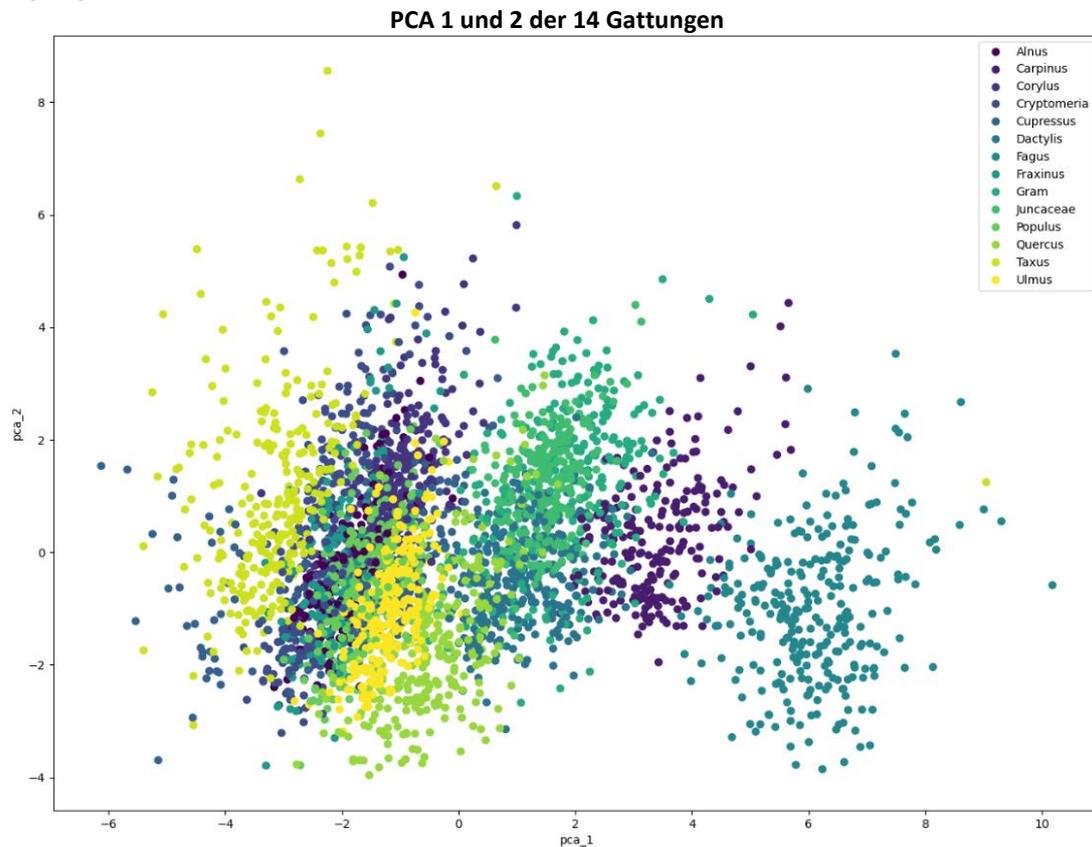


Abbildung 29: PCA Plot

Es ist zu sehen, dass sich einige Gattungen schon recht gut von den andern abgrenzen lassen. Diese Gattungen lassen sich später bei der Klassifikation der Gattungen wahrscheinlich schon recht gut klassifizieren. Einige andere Gattungen sind eher stark ineinander verwoben. Womöglich wird es bei diesen Gattungen bei der Klassifizierung eher Verwechslungen geben. Genaueres wird sich in der Klassifikation zeigen.

3.3.7 Fazit PCA

Mit PCA konnte nun der bisherige Datensatz mit den 18 originalen Features auf einen neuen Datensatz mit nur noch zehn PCA Features reduziert werden. Der neue Datensatz beinhaltet praktisch keine Redundanz zwischen den Features. Er beinhaltet noch fast 100 % der Varianz. Anhand der Eigenvektoren ist die Gewichtung der originalen Features für die neuen PCA Features bekannt. Eine Visualisierung der ersten beiden PCA Features gibt bereits ein erstes Gefühl für die nachfolgende Klassifikation der Gattungen.

3.4 Zusammenfassung

Für die folgende Klassifikation der Gattungen und die automatische Aussortierung sind die Daten aufbereitet. Es stehen jeweils ein Test- und ein Trainingsdatensatz zur Verfügung. Es ist jeweils ein Datensatz mit den originalen 18 Features sowie ein Datensatz mit den zehn PCA Features vorhanden. Bei der Datenstruktur handelt es sich jeweils um ein Pandas Dataframe. Der Aufbau ist im Kapitel 2.3.1 Aufbau Dataframe ersichtlich.

4 Klassifikation der Gattungen

In diesem Kapitel soll mittels SVM eine Klassifikation der Gattungen anhand der extrahierten Features vorgenommen werden. Zum einen soll eine Klassifizierung mit den originalen 18 Features erfolgen und zusätzlich eine mit dem neu transformierten Datenset mit den zehn PCA Features. Die beiden Klassifikationen sollen miteinander verglichen werden.

Es handelt sich somit um eine Mehrklassen-Klassifikation mit 14 unterschiedlichen Klassen. Bei den 14 unterschiedlichen Klassen handelt es sich um die 14 Pollengattungen, welche in der Einleitung kurz vorgestellt wurden. Es wird jeweils nur mit den guten Daten gearbeitet, also solchen, die ein Qualitätslabel von 1 haben.

4.1 Vorgehen

Der nachfolgende Ablauf beschreibt grob, wie die Klassifizierung realisiert wurde. Für das SVM Model wurde die Scikit-Learn Library verwendet.

Schritt 1: Import Trainingsdatenset

Als Erstes erfolgt der Import des Trainingsdatensets. Es werden nur die guten Daten importiert (`quality_lab = 1`).

Schritt 2: Vorverarbeitung Daten

Zuerst erfolgt das Extrahieren des Target Vektors und der Datenmatrix aus dem Datenset. Anschliessend muss die Datenmatrix vorverarbeitet werden. Bei der Variante ohne PCA beinhaltet das eine Standardisierung mit bereits bekannten Mittelwerten und Standardabweichungen (siehe Tabelle 12). Bei der Variante mit PCA kommt noch eine Datenmatrixtransformation mit der reduzierten Transformationsmatrix P (siehe Kap. 3.3.4) hinzu.

Schritt 3: Modell Fit

Anschliessend findet der Modell Fit statt. Hierzu werden der Targetvektor und die vorverarbeitete Datenmatrix benötigt.

Da bei einer SVM mehrere unterschiedliche Kernel-Funktionen und dazugehörige Parameter gewählt werden können, empfiehlt es sich, ein Hyperparameter tuning anzuwenden. Mittels Gridsearch und Crossvalidation können so die besten Parameter bestimmt werden. Das verwendete Grid ist in der Tabelle 16 ersichtlich.

Kernel-Typ	C	gamma	degree
linear	[0.001, 0.01, 0.1, 1, 10, 25, 50, 100, 1000]	-	-
rbf	[0.001, 0.01, 0.1, 1, 10, 25, 50, 100, 1000]	[1, 1e-1, 1e-2, 1e-3, 1e-4, 1e-5]	-
sigmoid	[0.001, 0.01, 0.1, 1, 10, 25, 50, 100, 1000]	[1, 1e-1, 1e-2, 1e-3, 1e-4, 1e-5]	-
poly	[0.001, 0.01, 0.1, 1, 10, 25, 50, 100, 1000]	[1, 1e-1, 1e-2, 1e-3, 1e-4, 1e-5]	[2, 3, 4,]

Tabelle 16: Parametergrid für das Hyperparameter tuning

Schritt 4: Validation mit Testdatenset

In einem letzten Teil kann mit dem Testdatenset das Modell validiert werden. Es ist zu vermerken, dass an den Testdaten jeweils die gleiche Vorverarbeitung vorgenommen werden muss, wie sie fürs Modellbild gemacht wurde.

4.2 Ergebnisse und anschließende Diskussion

Beim Ergebnisteil werden direkt beide Ergebnisse vorgestellt. Also jene, welche sich ohne PCA und mit PCA ergeben haben. Dies soll den Vergleich zwischen beiden vereinfachen. Anschliessend erfolgen direkt eine Interpretation und Diskussion der Ergebnisse.

Der erste Teil beinhaltet die Rangliste des Hyperparametertunings, und damit die Modellwahl. Diese wird mit dem Trainingsdatenset bestimmt.

Im zweiten Teil erfolgen dann die Ergebnisse der Validierung mit dem Testdatenset. Diese beinhalten jeweils eine Confusion Matrix sowie einen Klassifikationsreport.

4.2.1 Ranglisten Hyperparametertuning

Im Folgenden ist sowohl die Rangliste des Hyperparametertunings mittels Crossvalidation mit PCA als auch ohne PCA ersichtlich. Der 1. Rang wurde jeweils für das finale Modell verwendet. Beim Score handelt es sich um den Mittelwert der Accuracy von der Crossvalidation. Anschliessend wird mit der Abweichung die Standardabweichung des Scores angegeben. Danach werden der jeweilige dazugehörige Kernel-Typ und die dazugehörigen Parameter aufgeführt.

Ohne PCA					
Rang	Score	Abweichung	Kernel-Typ	Parameter C	Parameter gamma
1	0.726	(+/-0.016)	rbf	25	0.01
2	0.724	(+/-0.012)	rbf	50	0.01
3	0.724	(+/-0.020)	rbf	100	0.01
4	0.722	(+/-0.029)	linear	1000	kein gamma
5	0.722	(+/-0.022)	linear	100	kein gamma
6	0.72	(+/-0.028)	rbf	1	0.1

Tabella 17: Rangliste Parametertuning ohne PCA

Mit PCA					
Rang	Score	Abweichung	Kernel-Typ	Parameter C	Parameter gamma
1	0.717	(+/-0.031)	rbf	1	0.1
2	0.715	(+/-0.017)	rbf	25	0.01
3	0.714	(+/-0.024)	rbf	10	0.01
4	0.713	(+/-0.016)	rbf	50	0.01
5	0.712	(+/-0.021)	rbf	1000	0.001
6	0.712	(+/-0.021)	rbf	100	0.01

Tabella 18: Rangliste Parametertuning mit PCA

Die Unterschiede der Scores ohne und mit PCA sind ziemlich gering. Die Scores mit PCA fallen nur um ca. 1 Hundertstel schlechter aus als jene ohne PCA. Dies war auch schon zu erwarten, da mit der PCA ein nur sehr geringer Informationsverlust generiert und hauptsächlich Redundanz eliminiert wurde. Die Scores der ersten 6 Ränge sind jeweils sehr nahe beieinander. Die Differenz vom ersten zum sechsten Score liegt lediglich bei maximal 0.6 %.

Beim Kernel-Typ handelt es sich bis auf Rang 4 und 5 ohne PCA immer um einen rbf-Kernel. Dies bedeutet, dass sich die Gattungen im Vektorraum am besten durch eine nicht lineare Grenze abtrennen lassen.

Drei Kernel und die dazugehörigen Parameter aus der obigen Rangliste sind genau gleich wie bei der unteren.

4.2.2 Confusion Matrix und Klassifikationsbericht

Im Folgenden sind die beiden Confusion Matrizen, einmal ohne PCA und einmal mit PCA, ersichtlich. Bei den Spalten handelt es sich um die wahre Klasse und bei den Zeilen um die geschätzte Klasse. Die grau hervorgehobenen Daten in der Diagonalen entsprechen jeweils der korrekt klassifizierten Klasse. Alle anderen Einträge entsprechen den falsch klassifizierten Gattungen. Anhand der Confusion Matrix ist direkt ersichtlich, welche Gattung mit welcher verwechselt wurde.

		Actual class													
		Alnus	Carpinus	Corylus	Cryptomeria	Cupressus	Dactylis	Fagus	Fraxinus	Gram	Juncaceae	Populus	Quercus	Taxus	Ulmus
Predicted class	Alnus	103	0	9	33	0	0	0	33	0	0	22	4	0	9
	Carpinus	0	207	0	0	0	3	4	0	10	4	0	0	0	0
	Corylus	9	0	180	5	58	0	0	10	0	0	18	0	7	5
	Cryptomeria	53	0	2	171	0	0	0	23	0	0	11	1	11	0
	Cupressus	1	0	12	0	125	0	0	4	0	0	16	0	22	4
	Dactylis	0	7	0	0	0	212	0	0	68	32	0	13	0	0
	Fagus	0	2	0	0	0	0	288	0	0	0	0	0	0	0
	Fraxinus	34	0	16	18	3	0	0	91	0	0	11	0	6	3
	Gram	0	11	0	0	2	27	0	0	201	12	0	11	0	2
	Juncaceae	0	1	0	0	0	23	0	0	17	230	1	0	0	0
	Populus	42	0	12	7	29	3	0	15	0	0	182	0	1	33
	Quercus	0	0	0	0	0	23	0	0	4	0	0	255	0	1
	Taxus	2	0	4	3	33	0	0	2	0	0	2	0	215	0
	Ulmus	24	0	0	0	4	9	0	2	0	0	26	5	0	231

Tabelle 19: Confusion Matrix ohne PCA

		Actual class													
		Alnus	Carpinus	Corylus	Cryptomeria	Cupressus	Dactylis	Fagus	Fraxinus	Gram	Juncaceae	Populus	Quercus	Taxus	Ulmus
Predicted class	Alnus	102	0	8	36	0	0	0	31	0	0	26	1	0	6
	Carpinus	0	204	0	0	0	3	4	0	5	5	0	0	0	0
	Corylus	17	0	174	4	50	0	0	12	0	0	23	0	5	7
	Cryptomeria	55	0	4	172	0	0	0	25	0	0	18	2	7	2
	Cupressus	3	0	13	0	145	0	0	4	0	1	16	0	17	7
	Dactylis	0	8	0	0	0	197	0	0	66	30	0	16	0	0
	Fagus	0	3	0	0	0	0	287	0	0	0	0	0	0	0
	Fraxinus	34	0	12	17	1	0	0	92	0	0	9	1	5	2
	Gram	0	12	0	0	1	31	0	0	201	15	0	6	0	1
	Juncaceae	0	1	0	0	0	26	0	0	25	227	0	0	0	1
	Populus	35	0	14	3	25	5	0	10	0	0	165	0	1	36
	Quercus	0	0	0	0	0	30	0	0	3	0	0	257	0	1
	Taxus	1	0	8	5	30	0	1	3	0	0	2	0	227	0
	Ulmus	21	0	2	0	2	8	0	3	0	0	30	6	0	225

Tabelle 20: Confusion Matrix mit PCA

Die Seite beinhaltet die beiden Klassifikationsberichte. In diesen Berichten ist jeweils die Precision, der Recall und der f1-Score von jeder Klasse einzeln ersichtlich. Der Support-Wert gibt die Anzahl der Pollen einer Gattung im Testdatenset an. Zuerst wird jeweils zusammenfassend die Accuracy, der Macro Average und die Weighted Average angegeben.

Ohne PCA				
	precision	recall	f1-score	support
Alnus	0.48	0.38	0.43	268
Carpinus	0.91	0.91	0.91	228
Corylus	0.62	0.77	0.68	235
Cryptomeria	0.63	0.72	0.67	237
Cupressus	0.68	0.49	0.57	254
Dactylis	0.64	0.71	0.67	300
Fagus	0.99	0.99	0.99	288
Fraxinus	0.50	0.51	0.50	180
Gram	0.76	0.67	0.71	300
Juncaceae	0.85	0.83	0.84	278
Populus	0.56	0.63	0.59	289
Quercus	0.90	0.88	0.89	289
Taxus	0.82	0.82	0.82	262
Ulmus	0.77	0.80	0.78	288
accuracy			0.73	3700
macro avg	0.72	0.72	0.72	3700
weighted avg	0.73	0.73	0.73	3700

Tabelle 21: Klassifikationsbericht ohne PCA

Mit PCA				
	precision	recall	f1-score	support
Alnus	0.49	0.38	0.43	268
Carpinus	0.92	0.89	0.91	228
Corylus	0.60	0.74	0.66	235
Cryptomeria	0.60	0.73	0.66	237
Cupressus	0.70	0.57	0.63	254
Dactylis	0.62	0.66	0.64	300
Fagus	0.99	0.98	0.99	288
Fraxinus	0.53	0.51	0.52	180
Gram	0.75	0.67	0.71	300
Juncaceae	0.81	0.82	0.81	278
Populus	0.56	0.57	0.57	289
Quercus	0.88	0.89	0.89	289
Taxus	0.82	0.87	0.84	262
Ulmus	0.76	0.78	0.77	288
accuracy			0.72	3700
macro avg	0.72	0.72	0.72	3700
weighted avg	0.72	0.72	0.72	3700

Tabelle 22: Klassifikationsbericht mit PCA

Diskussion und Interpretation der Confusion Matrix und des Klassifikationsberichts

Als Erstes ist zu sagen, dass die Klassifizierungen ohne PCA und mit PCA fast gleich gut abschneiden. Das war anhand der fast gleichen Scores der Rangliste bei der Modellbildung auch zu erwarten. Die Accuracy ohne PCA beträgt 73 % und die Accuracy mit PCA 72 %. Die Differenz ist somit lediglich 1 %. Die Accuracy von nicht ganz 75 % bedeutet, dass von vier klassifizierten Gattungen drei im Durchschnitt korrekt klassifiziert sind. Dies ist jedoch, wie die weitere Erklärung zeigt, ziemlich abhängig von der jeweiligen Gattung. Einige Gattungen können schon mit einer grossen Wahrscheinlichkeit korrekt geschätzt werden. Andere Gattungen können nicht so gut voneinander unterschieden werden. Es kann dabei aber bereits gesagt werden, bei welcher anderen Gattung die Verwechslung am wahrscheinlichsten vorliegt. Es ist somit bereits eine Einschränkung möglich.

Da die Ergebnisse der Klassifikation ohne und mit PCA fast gleich herausgekommen sind, wird in der weiteren Diskussion nicht mehr zwischen beiden unterschieden. Es gilt für beide das Gleiche.

Anhand der Confusion Matrix ist gut ersichtlich, welche Gattung wie gut klassifiziert werden kann. Es ist auch direkt ersichtlich, zwischen welchen Gattungen öfters Verwechslungen vorliegen. Es ist zu erkennen, dass die Confusion Matrix nicht symmetrisch ist, aber leicht dazu neigt. Was bedeutet, wenn eine Polle der Gattung A oft mit einer Polle aus der Gattung B falsch klassifiziert wird, umgekehrt auch viele Pollen aus der Gattung B mit den Pollen aus der Gattung A falsch klassifiziert werden. Dies ist jedoch nicht immer der Fall. Zum Teil findet die Verwechslung auch nur einseitig statt.

Alnus, Cryptomeria, Fraxinus und Populus werden eher oft miteinander verwechselt. Corylus und Cupressus werden auch oft nicht korrekt voneinander unterschieden. Bei Dactylis werden einige falsch als Gram identifiziert. Diese sieben Gattungen sind auch jene, welche bei der Klassifizierung am schlechtesten abschneiden. Dies ist im Klassifikationsbericht gut ersichtlich. All diese Gattungen haben einen f1-Score kleiner als 70 %. Die weiteren sieben Gattungen haben jeweils einen f1-Score von grösser als 70 %. Fagus, Juncaceae und Quercus können schon ziemlich robust klassifiziert werden. Bei Fagus beträgt der f1-Score sogar 99 %.

Die Differenz zwischen Precision und Recall ist meistens eher gering. Das liegt daran, dass die Matrix, wie oben bereits erklärt wurde, leicht dazu neigt, symmetrisch zu sein. Die maximale Differenz beträgt ohne PCA 19 % bei Cupressus und mit PCA 14 % bei Corylus. Das ist dort, wo die Symmetrie am deutlichsten nicht vorhanden ist. Also finden auf der einen Seite viele Verwechslungen statt und auf der anderen nicht.

4.3 Abhängigkeit des Scores anhand der Anzahl PCA Features

Es hat sich gezeigt, dass die Klassifizierung mit dem reduzierten Datensatz fast gleich gut ausgefallen ist wie mit dem originalen Datensatz mit den 18 Features. Nun stellt sich die Frage, ob die Dimension des Datensatzes noch stärker reduziert werden kann und die Klassifizierung trotzdem gleich gut ausfällt.

Die Grafik zeigt die Scores, abhängig von der Anzahl PCA Features, welche für die Klassifizierung verwendet wurden. Die grüne Kurve zeigt die Scores des Trainingsdatensatzes und die blaue die Scores des Testdatensatzes.

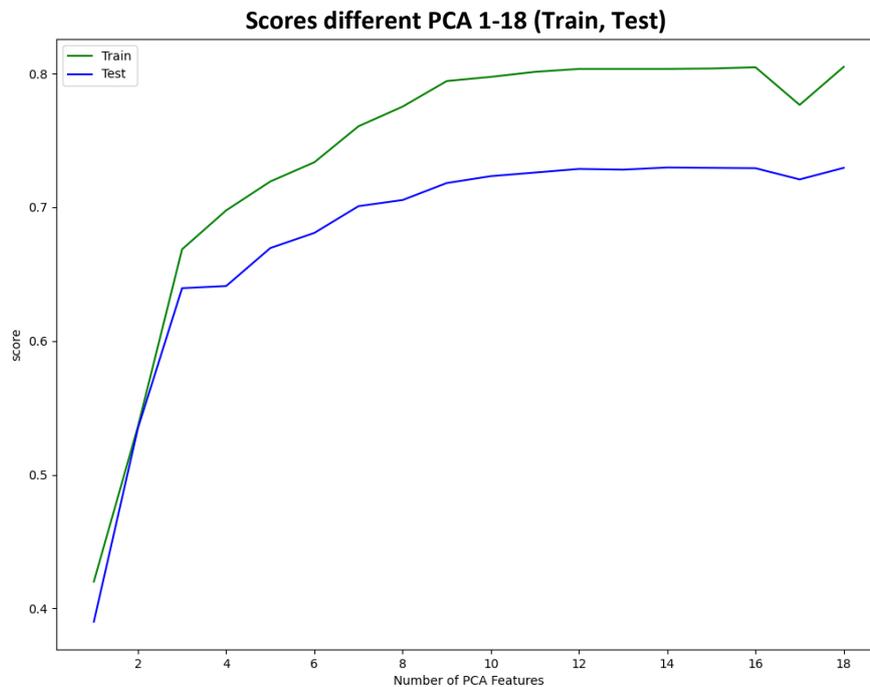


Abbildung 30: Scores PCA 1-18

Diskussion

Beide Kurven steigen bis zum dritten Feature stark an. Diese ersten drei Features sind somit für die Klassifizierung sehr wichtig. Bereits im PCA-Kapitel wurde in diesen drei Features der grösste Teil des Informationsgehaltes ausgemacht. Danach steigt die Kurve bis zum Feature zehn nur noch leicht an. Laut PCA nimmt der Informationsgehalt von PCA Feature vier bis zehn auch nur noch leicht zu. Ab Feature zehn flachen beide Kurven ziemlich stark ab, und es ist so gut wie fast keine Steigung mehr vorhanden. Mit Feature siebzehn sackt die Kurve sogar noch einmal ein. Wie im Kapitel PCA vorgestellt, liegt in den PCA-Features elf bis achtzehn auch kaum mehr ein Informationsgehalt. Die Kurve hat ziemlich Ähnlichkeit mit der Kurve der Abbildung 25 (siehe Kap. 3.3.3), was auch schon ein wenig zu erwarten war.

Somit ist die Feature-Zahl von zehn nicht schlecht gewählt. Bis dort steigt die Kurve noch einigermaßen leicht an. Danach ist fast keine Steigung mehr ersichtlich. Was überrascht, ist, dass mit nur drei Features bereits ein Testscore von ca. 63 % erreicht werden kann.

4.4 Fazit

Mit der SVM konnte sowohl eine Klassifikation der Gattungen mit dem Datensatz der 18 originalen Features als auch mit dem reduzierten Datensatz mit zehn PCA Features realisiert werden. Ein Vergleich der beiden Klassifikationen zeigte, dass beide ziemlich gleich gut sind. Mit einer Hyperparameter-Tuning konnten der beste Kernel und die dazugehörigen Parameter gefunden werden. Es wurde beim Testdatensatz eine Accuracy von fast 75 % erreicht.

Wie bereits im Vorfeld angenommen – zum einen bei der Betrachtung der Boxplot und damit der Variation zwischen den Arten und zum anderen bei der Visualisierung mittels PCA –, lassen sich einige Gattungen besser klassifizieren als andere. Anhand der Confusion Matrix konnte festgestellt werden, unter welchen Gattungen es zu Verwechslungen kommt. Es kann somit eingeschränkt werden.

In einem letzten Teil konnte noch untersucht werden, wie die Klassifikation ausfällt, abhängig von der Anzahl Features. Es konnte somit festgestellt werden, dass die neue Feature-Zahl von zehn PCA Features eine geeignete Wahl war.

Mit der SVM steht nun eine Alternative zu der bisherigen CNN-basierten Klassifikation zur Verfügung. Die Ergebnisse der SVM fallen nicht ganz so gut aus wie die des CNN-basierten Klassifikators. Die Gründe könnten folgende sein. Zum einen könnte es sein, dass die CNN-basierte Bildverarbeitung mehr Informationen aus den Bildern holen kann, wie zum Beispiel genauere Formeigenschaften. Zum anderen könnte das CNN noch tiefer auf Abhängigkeiten in und zwischen den Bildern eingehen.

Trotzdem konnte gezeigt werden, dass mit den extrahierten Features bereits eine Klassifikation realisiert werden kann. In einem weiteren Schritt wäre es vielleicht möglich die beiden Klassifikationen miteinander zu kombinieren.

5 Automatische Aussortierung

In den Datensets sind schlechte Bilder vorhanden (siehe Kap. 3.1). Eine manuelle Aussortierung der schlechten Bilder ist je nach Grösse des Datensets sehr zeitaufwendig. Zudem ist es nicht immer so einfach, zu entscheiden, ob es sich um ein schlechtes Bild handelt oder nicht, beispielsweise wenn es nicht so stark verschwommen ist.

Ziel ist es, eine solche Aussortierung von schlechten Bildern zu automatisieren. Es handelt sich somit um eine binäre Klassifizierung, die jeweils entscheiden muss, ob es sich um gute oder schlechte Qualität handelt. Das Konzept sieht vor, den bereits vorhandenen Klassifikator der Gattungen dafür zu verwenden. Anhand des Wahrscheinlichkeitswertes für eine bestimmte Gattung kann entschieden werden, ob es sich bei den Daten im Datenset um eine gute oder schlechte Qualität handelt. Da die Klassifikation der Gattungen mit und ohne PCA fast gleich gut herausgekommen ist, wird nur noch mit einem Modell weitergearbeitet, und zwar mit PCA.

5.1 Ablauf Programm

Der Ablauf der automatischen Aussortierung sieht wie folgt aus:

Input

Als Input ist ein gelabeltes Datenset erforderlich. Es muss bekannt sein, um welche Gattung es sich handelt. Es muss sich um eine der 14 bekannten Gattungen handeln.

Schritt 1: Datenvorverarbeitung

Die Daten müssen als Erstes wieder gleich vorverarbeitet werden wie beim Modell Fit. Da hier nun mit dem Modell PCA gearbeitet wird, bedeutet dies als Erstes eine Standardisierung mit den bereits bekannten Mittelwerten und Standardabweichungen (siehe Kap. 3.2.4) und eine anschliessende Basistransformation mit der reduzierten Transformationsmatrix P (siehe Kap. 3.3.4).

Schritt 2: Wahrscheinlichkeit Vorhersage

Anschliessend erfolgt mit dem trainierten Modell eine Wahrscheinlichkeitsschätzung für die gelabelte Gattung.

Schritt 3: Labeln anhand Schwellwert

Anhand eines eingestellten Schwellwerts, welcher für jede Gattung separat eingestellt werden kann, erfolgt eine Zuweisung des Labels für Qualität (`quality_pred`). Ist die Wahrscheinlichkeit grösser als der Schwellwert, wird dem Label die Zahl 1 zugeteilt. Diese bedeutet, dass es sich um eine gute Qualität handelt (siehe Kapitel 2.3.1). Ist der Wahrscheinlichkeitswert kleiner als der Schwellwert, wird dem Label die Zahl -1 zugeteilt. Es wird somit als qualitativ schlecht deklariert.

<code>quality_pred = 1</code> , wenn grösser als der Schwellwert	Gute Qualität geschätzt
<code>quality_pred = -1</code> , wenn kleiner als der Schwellwert	Schlechte Qualität geschätzt

Schritt 4: Sortierung mit geschätztem Label

Anhand des geschätzten Qualitätslabels kann nun das Datenset nach guter und schlechter Qualität sortiert werden. Die schlecht gelabelten Bilder werden in einen Ordner BAD und die gut gelabelten Bilder in einen Ordner GOOD verschoben. In der aktuellen Version ist es erforderlich, dass sich bereits ein leerer Ordner GOOD sowie BAD im Verzeichnis des Datensets befinden.

<code>Quality_pred = 1</code>	Verschiebung von <code>Im0</code> , <code>Im1</code> und des JSON-Files in den Ordner GOOD
<code>Quality_pred = -1</code>	Verschiebung von <code>Im0</code> , <code>Im1</code> und des JSON-Files in den Ordner BAD

5.2 Ergebnisse und anschließende Diskussion

Für die Validierung wird im ersten Teil ein Histogramm und im zweiten Teil eine ROC-Kurve verwendet. Die Präsentation der Ergebnisse erfolgt mit jeweils vier repräsentativen Gattungen. Das wären: Juncaceae, Cryptomeria, Alnus und Populus. Juncaceae repräsentiert die am besten und Populus die am schlechtesten abgeschnittene Gattung. Die anderen beiden repräsentieren das Mittelfeld. Die restlichen Grafiken zu allen Gattungen sind im Anhang zu finden.

5.2.1 Histogramme der Wahrscheinlichkeitsschätzung

In den Abbildungen 31 und 32 sind die absoluten Häufigkeiten der Wahrscheinlichkeitsschätzung ersichtlich. Auf der X-Achse ist die Wahrscheinlichkeitsschätzung des Klassifizierers für die jeweilige Gattung ersichtlich. Die Klassenbreite beträgt jeweils 5 %. Auf der Y-Achse sind die absoluten Häufigkeiten eingetragen.

Auf der linken Seite sind die Ergebnisse vom Trainingsdatenset ersichtlich und auf der rechten jene des Testdatensets. In Grün eingefärbt sind jeweils die gut gelabelten und in Rot die schlecht gelabelten Daten.

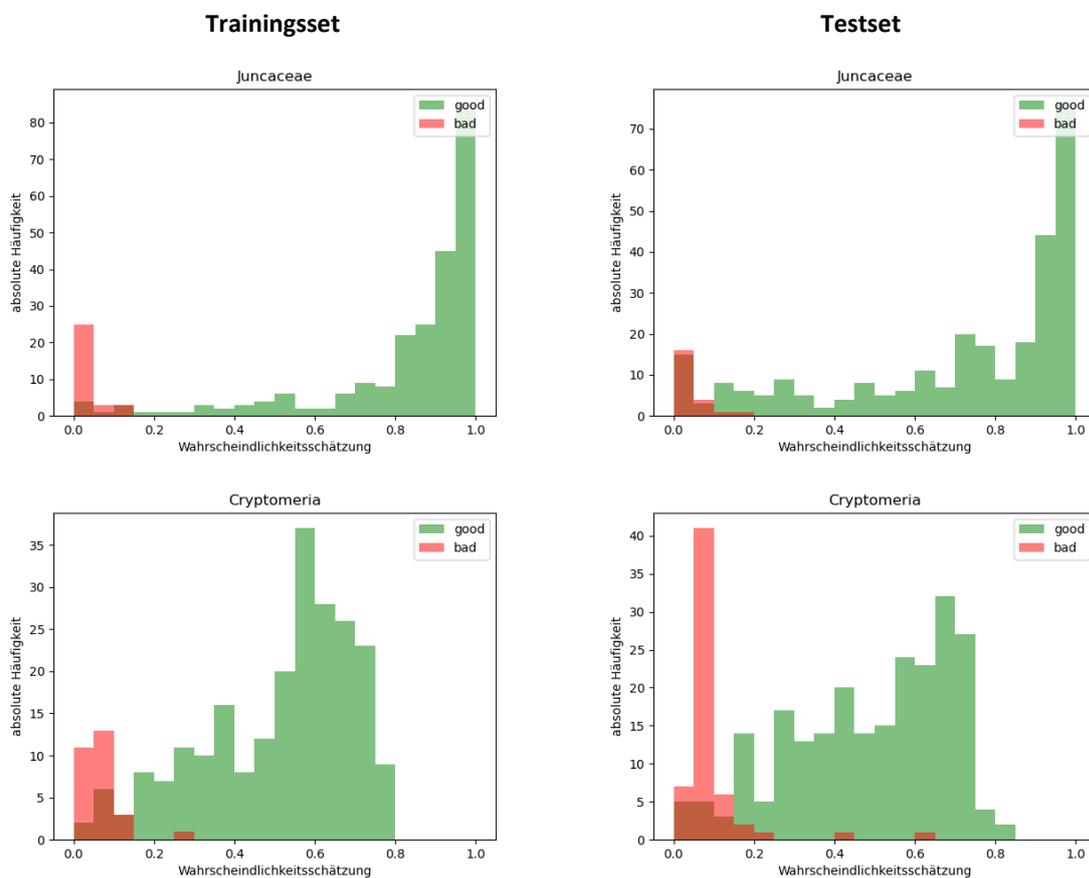


Abbildung 31: Histogramme der Wahrscheinlichkeitsschätzung Teil 1

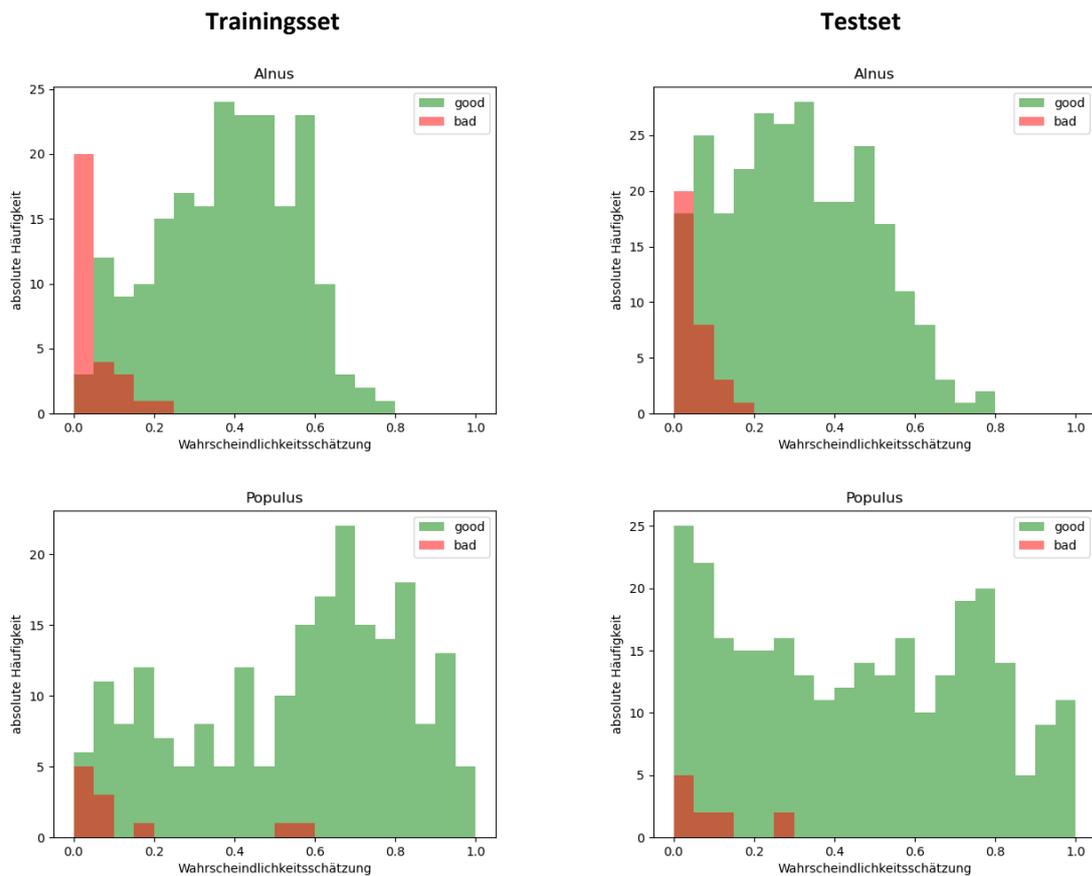


Abbildung 32: Histogramme der Wahrscheinlichkeitsschätzung Teil 2

Diskussion

Wie im Theorieteil bereits erklärt, ist anhand dieser Grafiken der Einfluss des Schwellwertes auf die Klassifizierung gut ersichtlich.

Es ist zu erkennen, dass die Ergebnisse aus dem Test- und Trainingsdatensatz ziemlich ähnlich ausgefallen sind. Die Ergebnisse aus dem Testdatensatz sind nur wenig schlechter. Das war auch anzunehmen, da die Daten des Trainingssets auch für die Modellbildung verwendet wurden. Der nur kleine Unterschied spricht für das Modell.

In der Grafik ist gut zu erkennen, dass es sich um Imbalanced Informationen handelt. So ist der Anteil an schlechten Daten um einiges kleiner als der Anteil an guten Daten. Im Durchschnitt beträgt das Verhältnis ca. 5 bis 10 %. Dies ist jedoch sehr gattungsabhängig. Es gibt Gattungen, die fast keine schlechten Bilder enthalten.

Das Ergebnis für Juncaceae fällt sehr gut aus. Die Wahrscheinlichkeitsschätzungen für die guten Bilder fallen alle sehr hoch aus, also nahe bei eins. Die Wahrscheinlichkeiten für die schlechten Bilder kommen kaum über 15 %. Es ist nur ein geringer Anteil der guten Bilder unter 20 %. Juncaceae ist auch eine Gattung, welche bei der Klassifizierung der Gattungen bereits sehr gut abgeschnitten hat.

Die Wahrscheinlichkeitsschätzungen für *Cryptomeria* und *Alnus* für die guten Bilder fallen nicht so hoch aus. Sie sind also nicht so nahe bei eins. Der Grund ist auch wieder, wie bereits in der Klassifizierung der Gattungen ersichtlich war, dass diese Gattungen häufiger mit anderen Gattungen verwechselt werden.

Bei den meisten Gattungen ist zu erkennen, dass der Wahrscheinlichkeitsschätzwert bei den schlechten Daten meistens nicht über 20 % kommt. Dies könnte somit grob ein Schwellwert sein, bei welchem nur sehr wenig schlechte Daten als gut deklariert werden. Der Schwellwert kann aber bei jeder Gattung einzeln bestimmt werden.

5.2.2 ROC-Kurven

In den Grafiken sind die ROC-Kurven der vier repräsentativen Gattungen ersichtlich. In Grün ist jeweils die Kurve des Trainingssets und in Orange die Kurve des Testsets eingezeichnet. Die blau gestrichelte Linie dient nur zur besseren Übersicht und hat keine spezielle Bedeutung. Auf der X-Achse befindet sich jeweils die False Positive Rate und auf der Y-Achse die True Positive Rate (siehe Kapitel 2.6.3).

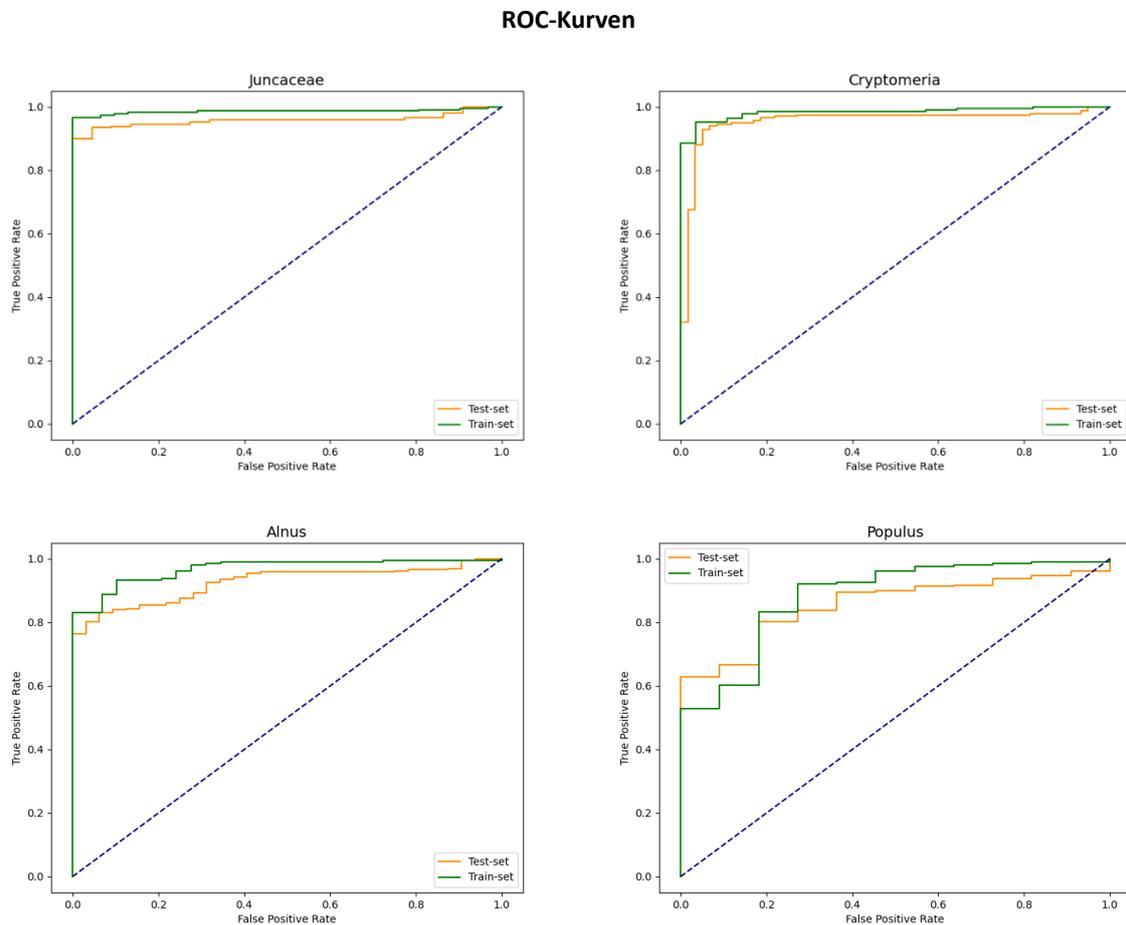


Abbildung 33: ROC-Kurven

Diskussion

Grundsätzlich ist in den Grafiken dieselbe Information wie in den obigen Histogrammen enthalten. In diesen Grafiken sind jedoch direkt die beiden Raten, die zu erreichen sind, ersichtlich.

Bei Juncaceae zeigen die beiden Kurven einen fast idealen Verlauf. Sie erreichen fast den Punkt (0,1). Es ist somit eine sehr hohe TPR bei einer gleichzeitig sehr geringen FPR zu erreichen. Die Kurvenverläufe von Cryptomeria und Alnus sind auch noch ziemlich gut. Bei Populus beginnt die FPR bereits bei einer TPR von ca. 60 % zuzunehmen. Will man somit die FPR möglichst tief halten, bedeutet dies, dass viele gute Bilder als schlecht klassifiziert werden (FN). Das wurde bereits vorhin auch schon festgestellt.

5.3 Fazit

Es konnte eine Aussortierung der schlechten Bilder mit dem bereits trainierten Klassifikator für die Gattungen erstellt werden. Es musste lediglich eine Wahrscheinlichkeitsschätzung zum bereits bestehenden SVM-Modell hinzugefügt werden. Das Programm, welches die automatische Aussortierung erledigt, konnte bereits implementiert werden.

Für die automatische Aussortierung ist für jede Gattung einzeln ein Schwellwert zu bestimmen. Der Schwellwert kann mithilfe der erstellten Histogramme und ROC-Kurven nun nach den Bedürfnissen gewählt werden.

Die Aussortierung funktioniert bei den 14 Gattungen schon recht gut. Die meisten schlechten Daten können gut gefunden werden. Je nach Gattung kommt es zu mehr oder weniger FN, also eigentlich guten Daten, die als schlecht klassifiziert werden.

6 Schluss

Der Schluss beinhaltet als Erstes ein Fazit der Gesamtarbeit. Hier wird kurz zusammengefasst, was in der Arbeit bereits erreicht wurde und welche Punkte noch nicht angegangen oder erledigt werden konnten. Mit einem Ausblick wird dann auf jene Punkte eingegangen, welche man als nächste bearbeiten könnte. Abschliessend erfolgt ein kurzes persönliches Schlusswort.

6.1 Fazit

Mit der Pollenrecherche konnten die Merkmale gefunden werden, welche für die Beschreibung einer Polle üblich sind. Es wurden dabei jene Merkmale herausgesucht, welche mittels Bildverarbeitung (an den zwei zur Verfügung stehenden Bildern) herausgeholt werden können. Es konnte somit eine Merkmalliste erstellt werden mit den aus der Biologie üblichen Merkmalen für die Beschreibung einer Polle. Die bisherige Bildverarbeitung wurde neu mit der OpenCV-Library implementiert. Es sind einige Features hinzugefügt worden, welche zum einen für die Anpassung an die «neue» Merkmalliste oder zum andern für die Kontrolle der Qualität der Bildpaare nützlich sein könnten. Die Umrechnungen von digitalen auf die physikalischen Einheiten konnten bereits realisiert werden. Mit dem Pandas Dataframe konnte eine Datenstruktur erstellt werden, die das Arbeiten mit den grossen Datenmengen sehr vereinfacht. Die Datenstruktur ist ausgerichtet auf ein Filehandling, Klassifikationsaufgaben und die Datenanalyse im Allgemeinen.

Mit einer ersten Datenanalyse an den zur Verfügung stehenden Daten konnten unterschiedliche Fehlerquellen und ihre Auswirkungen aufgefunden gemacht werden. Es war auch möglich, bereits eine erste Begutachtung der Variation zwischen den Arten vorzunehmen. Zudem wurden für die weiterführende Arbeit ein Trainings- und ein Testdatenset erstellt. Diese beinhalten zwei unterschiedliche Labels, einerseits das Label der Gattung und andererseits jenes der Qualität. Mit den guten Daten aus dem Trainingsdatenset konnte eine tiefere Analyse der Daten durchgeführt werden. Diese gab genaueren Einblick in die Variation innerhalb einer Klasse und in die Abhängigkeiten der Features. Eine erste 3D-Schätzung konnte bereits vorgenommen werden. Für eine automatische 3D-Form-Bestimmung ist jedoch leider noch kein Konzept gefunden worden. Es konnte ermittelt werden, welche Features für die Klassifikation der Gattungen verwendet werden und wie diese genau aufgebaut sind. Mittels PCA war es möglich, die Dimension des Datensets von 18 auf 10 zu reduzieren, ohne einen grossen Informationsverlust zu generieren. Es wurde dabei hauptsächlich Redundanz zwischen den Features eliminiert. Die Aufbereitung der Daten stellt somit zwei unterschiedliche Datensets für die weiterführende Arbeit zusammen: das Datenset mit den originalen 18 Features und das mittels PCA reduzierte Datenset mit 10 PCA Features.

Mit einer SVM konnte eine Klassifikation der Gattungen mithilfe der extrahierten Features realisiert werden. Somit steht eine Alternative zu der bisherigen CNN-basierten Klassifikation zur Verfügung. Ein Vergleich hat gezeigt, dass die Klassifikation mit dem Datensatz mit den 10 PCA Features gleich gut funktioniert wie mit den 18 originalen Features. Mit dem Klassifikator der Gattungen konnte gleichzeitig eine automatische Aussortierung eines gelabelten Datensets entwickelt werden. Die Aussortierung erfolgt anhand eines Schwellwertes. Dieser kann mithilfe von erstellten Histogrammen und ROC-Kurven für jede Gattung nach den Bedürfnissen eingestellt werden. Somit können die Datensets automatisch und nicht mehr mühsam und aufwendig manuell gesäubert werden. Es konnten somit schon einige Punkte der Aufgabenstellung erledigt werden. Für einige Bereiche hat die Zeit jedoch nicht ausgereicht.

Es konnte noch keine automatische 3D-Form-Schätzung der Polle entwickelt werden. Erste Untersuchungen haben gezeigt, dass eine solche 3D-Form-Schätzung nicht ganz so einfach ist und noch einmal einiges an Zeit beanspruchen würde. Erst mit einer 3D-Form-Schätzung kann die Feature-Extraktion an die «neue» Merkmalliste angepasst werden. Für die «neue» Merkmalliste müssen nämlich die Polar- und die Äquatorialansicht bekannt sein. Mit diesen können dann die restlichen Features abgeleitet werden. Die Zusatzanforderungen konnten auch noch nicht bearbeitet werden. Das waren einerseits die Verbesserung der Feature-Extraktion bei kleinen Partikeln und andererseits die Feature-Extraktion speziell für Sporen. Grund dafür war, dass mit der automatischen Aussortierung der schlechten Bilder während der Bearbeitung eine

neue Aufgabe dazugekommen ist. Die Bearbeitung dieses Problems wurde höher priorisiert und dementsprechend zuerst bearbeitet.

6.2 Ausblick

Im Ausblick wird kurz auf einige Punkte eingegangen, die als nächste bearbeitet werden könnten. Bei den ersten beiden Punkten handelt es sich um mögliche Erweiterungen der bereits bearbeiteten. Im letzten Punkt wird noch kurz auf die Feature-Extraktion eingegangen, speziell auf die 3D-Form-Schätzung.

6.2.1 Betreff Klassifikation der Gattungen

Die bisherige Klassifikation der Gattungen wurde mit einer SVM realisiert. Es wäre interessant herauszufinden, ob mit einem anderen Klassifikator, wie zum Beispiel mit einem Multilayer Perceptron (MLP), noch mehr herausgeholt werden könnte. Zudem könnte noch untersucht werden, ob die Klassifikation mit zusätzlichen Features, wie zum Beispiel mit der Orientierung, noch verbessert werden könnte, oder ob eine spätere 3D-Form-Schätzung zu einer Verbesserung der Klassifikation führen würde. Die bisherige Klassifikation bietet somit eine gute Möglichkeit für Vergleiche.

6.2.2 Automatische Aussortierung

Die bisherige automatische Aussortierung weist zwei Nachteile auf. Zum einen funktioniert sie nur auf gelabelten Datensätzen, und es muss sich um eine der 14 bekannten Gattungen handeln. Zum anderen ist die automatische Aussortierung nur eine binäre Klassifikation, also entweder gut oder schlecht.

Es könnte eine automatische Aussortierung entwickelt werden, die ohne Vorkenntnisse der Gattung bereits eine Qualitätsaussage über die beiden Bilder macht. Zudem könnte die Klassenzahl vergrößert werden. Sie könnte die unterschiedlichen Arten von schlechten Bildern (siehe Kap. 3.1.3) direkt anhand gewisser Merkmale als solche klassifizieren.

6.2.3 Feature-Extraktion

In einem ersten Schritt sollte getestet und validiert werden, ob die neue Feature-Extraktion mit OpenCV qualitativ die gleich guten oder besseren Resultate liefert wie die bisherige. Zudem sollte untersucht werden, ob die gewünschten Performance-Verbesserungen auch wirklich vorhanden sind.

Im Weiteren soll versucht werden, eine 3D-Form-Schätzung zu realisieren. In einem ersten Schritt würde es sich empfehlen, mit den zwei Ellipsen und der Orientierung ein Rotationsellipsoid zu schätzen, wie das manuell in Kapitel 3.2.2 versucht wurde. Es soll ein Rotationsellipsoid geschätzt werden, da es sich um eine sehr einfache Form handelt. Erst nach einer erfolgreichen Schätzung eines solchen Körpers empfiehlt es sich, eventuell noch weitere Formen abzuschätzen.

Aus den bisherigen Überlegungen sind mögliche zwei Ansätze entstanden, mit welchen ein Rotationsellipsoid geschätzt werden könnte. Dies sind lediglich zwei grobe Ideen. Es gibt sicherlich noch weitere.

- 1) Mathematischer Ansatz

Vielleicht existieren bereits mathematische Formeln oder es kann mathematisch hergeleitet werden.

- 2) Ansatz mit Netzwerk trainieren

Mit gelabelten Daten könnte ein Netzwerk trainiert werden, welches die gewünschten Daten schätzt.

Mit einer 3D-Form-Schätzung könnten anschliessend eine Polar- und Äquatorialansicht definiert werden. Mit diesen beiden Ansichten lassen sich die restlichen Merkmale aus der «neuen» Merkmalliste ableiten. Somit könnte die Feature-Extraktion an die «neue» Merkmalliste angepasst werden.

6.3 Schlusswort

Die Arbeit war aus meiner Sicht sehr interessant. Es hat mich immer wieder fasziniert, wie mittels der Holografie eine solche Auflösung an freischwebenden Teilchen möglich ist. Das Maschinelle Lernen ermöglicht interessante Anwendungen und ist ein sehr mächtiges Tool. Ich war überrascht, was die Statistik dabei für eine zentrale Rolle spielt.

Die Aufgaben waren für mich jedoch sehr anspruchsvoll und zeitintensiv. Speziell der Umgang mit den grossen Datenmengen war am Anfang eine grosse Hürde. Zum einen waren es die Anzahl an Klassen, also 14, und zum andern die Dimensionalität von teils grösser als 20, welche mich zu Beginn etwas überforderten. Mit dem Einsatz einer geeigneten Datenstruktur und mit dem Kennenlernen gewisser Operationen konnte der Umgang mit den grossen Datenmengen jedoch gemeistert werden. Es hat sich gezeigt, dass eine Analyse mit solch grossen Daten sehr schnell sehr aufwendig werden kann und dass der Überblick schnell verloren geht. Methoden des Maschinellen Lernens werden dabei fast unausweichlich.

Durch die Hürden, die es zu überwinden gab, konnte ich aber auch dementsprechend viel in diesem Bereich lernen. Vor allem in Bezug auf das Arbeiten mit Python, speziell in der Bildverarbeitung, der Datenanalyse und des Maschinellen Lernens, kann ich rückblickend sagen, dass ich mir enorm viel neues Wissen aneignen konnte. Es gelang mir, einige Ergebnisse zu generieren, die hoffentlich auch für den Industriepartner von Nutzen sein werden.

Als Erstes möchte ich mich bei der Swisens AG bedanken, welche diese Arbeit erst ermöglichte. Auch bedanke ich mich für die aus meiner Sicht gute Zusammenarbeit mit Herrn Prof. Dr. Zahn als Betreuer der Arbeit und bei Herrn Zeder als Ansprechperson des Industriepartners. Bei Fragen und Unklarheiten haben Sie sich immer Zeit genommen. Probleme konnten gemeinsam diskutiert werden, und es wurden mir stets gute Tipps für die Bearbeitung gegeben. Besten Dank dafür.

Abbildungsverzeichnis

Abbildung 1: Schematischer Aufbau Swisens Poleno	1
Abbildung 2: Messaufbau Holografie (Bächler, 2017)	2
Abbildung 3: Holografie-Funktionsweise	2
Abbildung 4: Feature-Extraktion.....	3
Abbildung 5: Apertur und Oberflächenbeschaffenheit (AutPal, 2020).....	5
Abbildung 6: Mögliche Polleneinheiten (Huffner)	6
Abbildung 7: Polar- und Äquatorialansichten (Huffner).....	6
Abbildung 8: Beispielbild Input.....	8
Abbildung 9: Finden des Thresholds.....	8
Abbildung 10: Erstellung des Binary-Bildes.....	9
Abbildung 11: Labeln des Binary Image.....	9
Abbildung 12: Region-Properties-Berechnung	9
Abbildung 13: Neue Merkmale	10
Abbildung 14: SVM (Chakure, 2020).....	14
Abbildung 15: Abhängigkeit eines Schwellwertes mittels Verteilung (roc-curves-Wikipedia, 2020).....	16
Abbildung 16: ROC-Kurve (roc-curves-Wikipedia, 2020).....	16
Abbildung 17: Boxplot der Fläche	18
Abbildung 18: Klassen von schlechten Bildern	19
Abbildung 19: Ellipsen-Fit.....	21
Abbildung 20: Definition von Image 0 und Image 1.....	22
Abbildung 21: Histogramm und Scatterplot.....	23
Abbildung 22: Boxplot-Orientierung	25
Abbildung 23: Kovarianzmatrix von X mit Features als Hilfe	28
Abbildung 24: Transformationsmatrix P und Eigenwerte.....	29
Abbildung 25: Abhängigkeit der Varianz anhand der Anzahl PCA Features	29
Abbildung 26: PCA 10.....	30
Abbildung 27: Transformation des Datensatzes.....	31
Abbildung 28: Kovarianzmatrix von Y.....	31
Abbildung 29: PCA Plot	32
Abbildung 30: Scores PCA 1-18	38
Abbildung 31: Histogramme der Wahrscheinlichkeitsschätzung Teil 1	41
Abbildung 32: Histogramme der Wahrscheinlichkeitsschätzung Teil 2	42
Abbildung 33: ROC-Kurven	43

Tabellenverzeichnis

Tabelle 1: Extrahierte Features.....	3
Tabelle 2: Liste der Pollengattungen.....	4
Tabelle 3: Grössenunterteilung Pollen (Heidemarie Halbritter, 2018).....	6
Tabelle 4: PE-Verhältnisse (Erdtman, 1986).....	7
Tabelle 5: Aufbau Dataframe	11
Tabelle 6: Codierung Qualität	11
Tabelle 7: Notation PCA	12
Tabelle 8: Kernel-Typen und Parameter	14
Tabelle 9: Mögliche Ergebnisse bei einer Klassifikation	15
Tabelle 10: Confusion-Matrix-Beispiel.....	15
Tabelle 11: 3D-Form-Schätzung	21
Tabelle 12: Mittelwerte und Standardabweichungen der Features.....	24
Tabelle 13: Struktur Datensatz mit 18 Features	27
Tabelle 14: Notation PCA	27
Tabelle 15: Die ersten drei PCA Features	30
Tabelle 16: Parametergrid für das Hyperparametertuning	33
Tabelle 17: Rangliste Parametertuning ohne PCA.....	34
Tabelle 18: Rangliste Parametertuning mit PCA	34
Tabelle 19: Confusion Matrix ohne PCA.....	35
Tabelle 20: Confusion Matrix mit PCA	35
Tabelle 21: Klassifikationsbericht ohne PCA	36
Tabelle 22: Klassifikationsbericht mit PCA.....	36

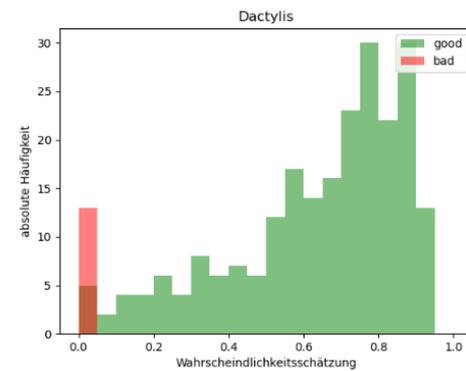
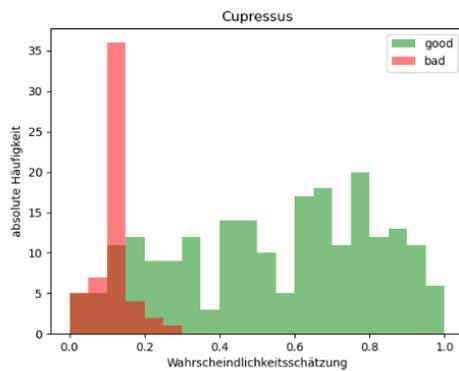
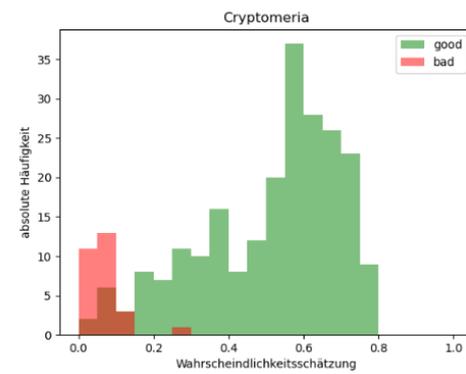
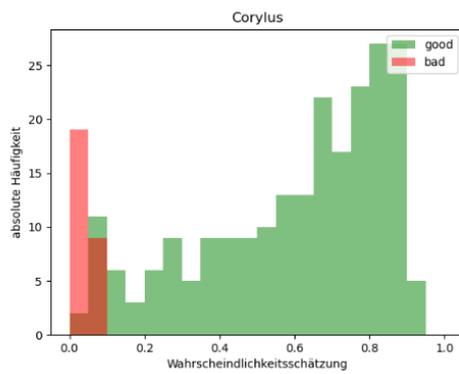
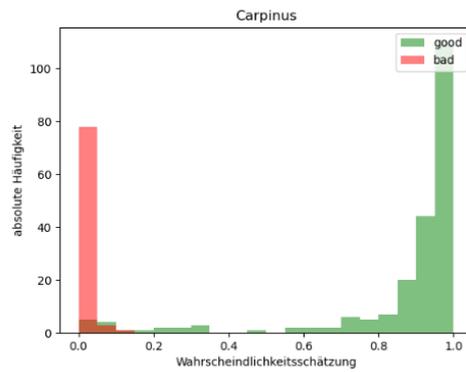
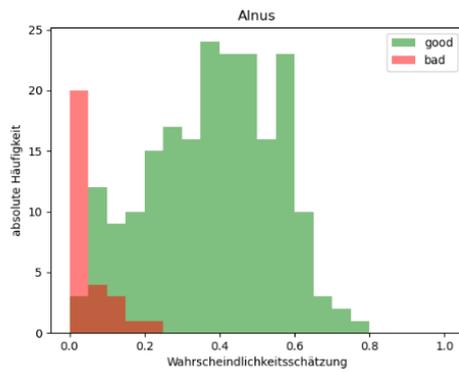
Literaturverzeichnis

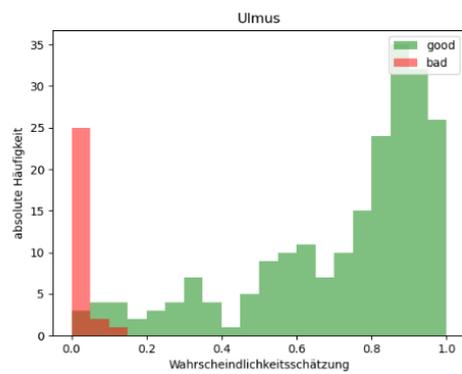
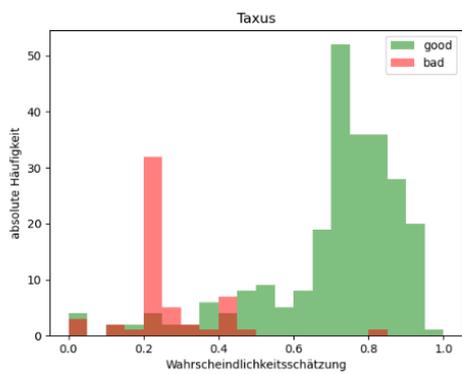
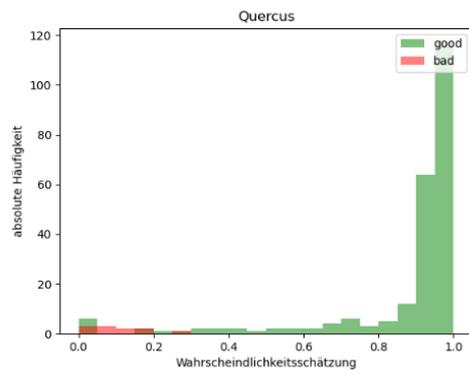
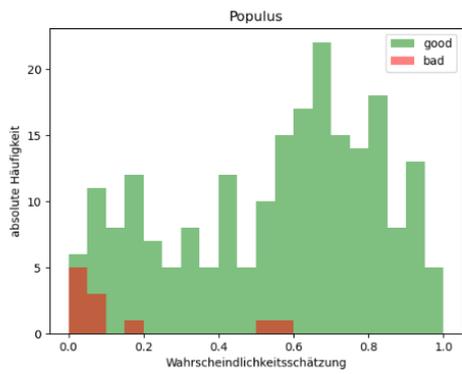
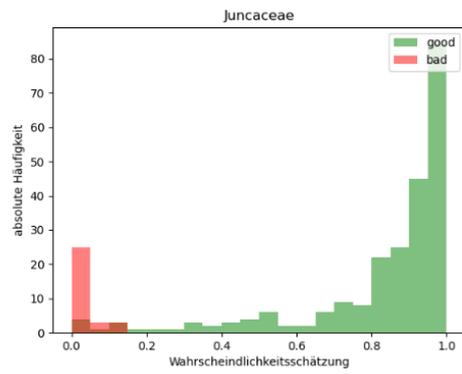
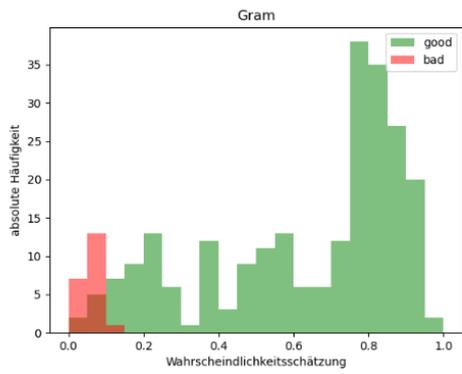
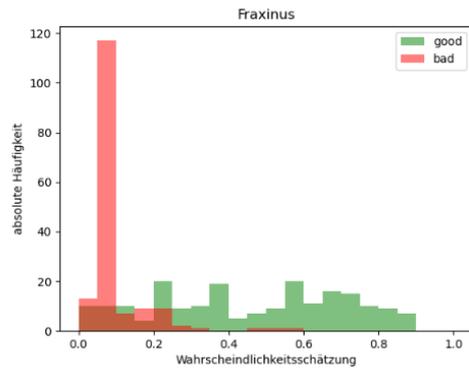
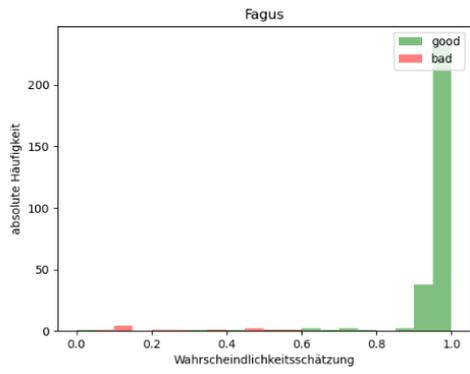
- AutPal. (21. Dezember 2020). *paldat*. Von <https://www.paldat.org/> abgerufen
- Bächler, P. (2017). *Paind Arbeit: Partikelerkennung mittels Holographie*. Alpnach.
- Chakure, A. (12. 12 2020). *aaaanchakure*. Von Support Vector Machines: <https://aaaanchakure.medium.com/support-vector-machines-svms-4bccbd78369> abgerufen
- Erdtman, G. (1986). *Pollen Morphology and Plant Taxonomy*. Stockholm: Hafner Publishing Company.
- Gaussianer. (2. 11 2020). *Wikipedia*. Von <https://de.wikipedia.org/wiki/Holografie> abgerufen
- Haß, J. (11. 12 2020). *Wikipedia*. Von Beurteilung eines binären Klassifikators: https://de.wikipedia.org/wiki/Beurteilung_eines_bin%C3%A4ren_Klassifikators abgerufen
- Heidemarie Halbritter, S. U.-R. (2018). *Illustrated Pollen Terminology*. Springer Open.
- Hossen, M. A. (21. 12 2020). *towardsdatascience*. Von Top Python Libraries: Numpy & Pandas: <https://towardsdatascience.com/top-python-libraries-numpy-pandas-8299b567d955> abgerufen
- Huffner, D. (kein Datum). Quick Reference Glossary with Illustrations.
- Myriantous, G. (23. 12 2020). *towards data science*. Von Feature Scaling and Normalisation in a nutshell: <https://towardsdatascience.com/feature-scaling-and-normalisation-in-a-nutshell-5319af86f89b> abgerufen
- OpenCV-team. (12. 10 2020). *opencv-about*. Von <https://opencv.org/about/> abgerufen
- Pandey, P. (20. 12 2020). *opensource*. Von 10 Python image manipulation tools: <https://opensource.com/article/19/3/python-image-manipulation-tools> abgerufen
- roc-curves-Wikipedia. (12. 12 2020). *Wikipedia*. Von Receiver operating characteristic: https://en.wikipedia.org/wiki/Receiver_operating_characteristic abgerufen
- scikit-image-development-team. (12. 10 2020). *scikit-image image processing in python*. Von <https://scikit-image.org/> abgerufen
- scikit-learn-authors. (16. 12 2020). *scikit-learn*. Von Support Vector Machines: <https://scikit-learn.org/stable/modules/svm.html> abgerufen
- Shlens, J. (25. 3 2003). A TUTORIAL ON PRINCIPAL COMPONENT ANALYSIS, Derivation, Discussion and Singular Value Decomposition.
- Swisens. (17. 12 2020). *swisens*. Von swisens.ch abgerufen
- Wikipedia. (23. 12 2020). *Kovarianzmatrix*. Von <https://de.wikipedia.org/wiki/Kovarianzmatrix> abgerufen

Anhang

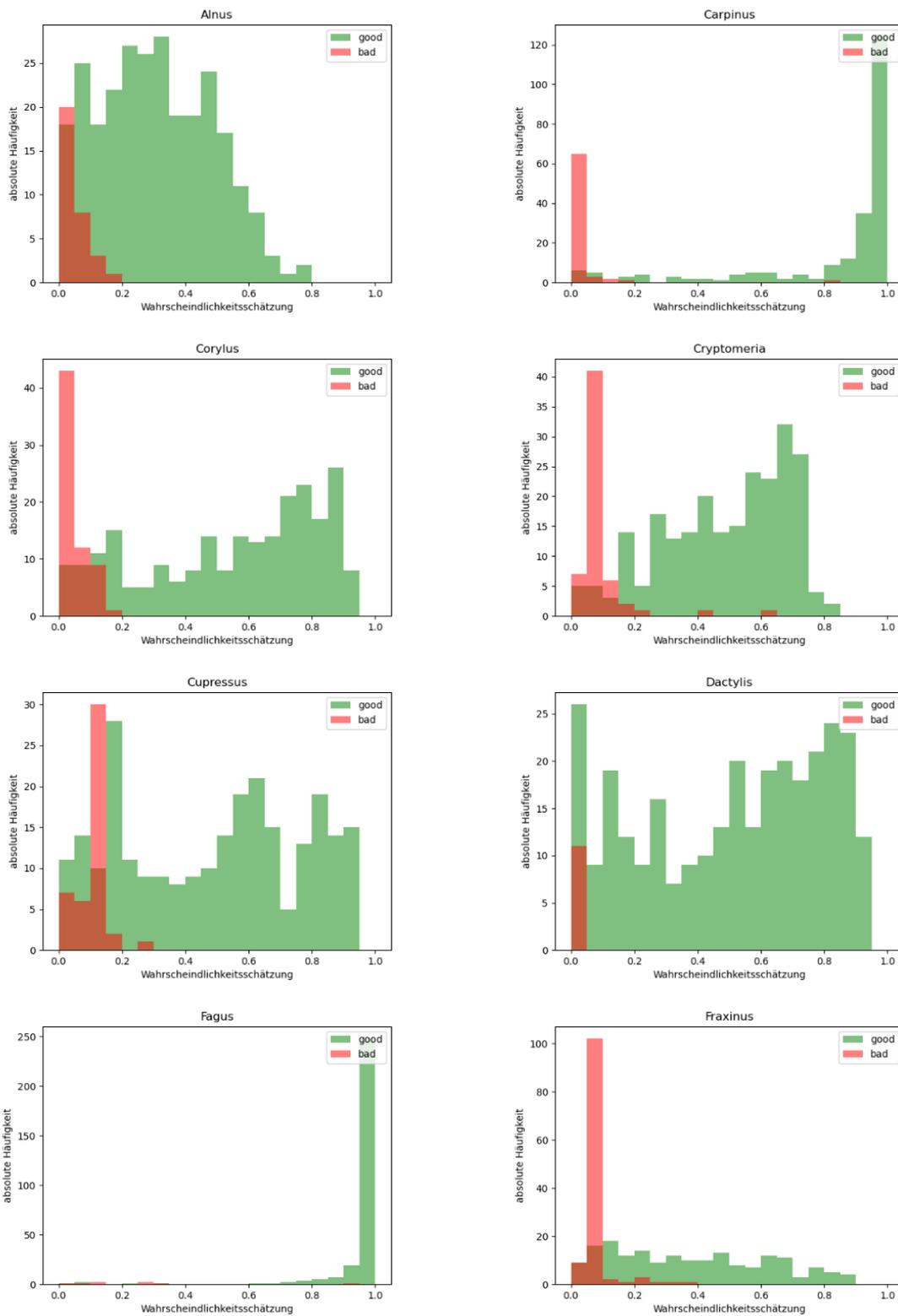
Grafiken automatische Aussortierung

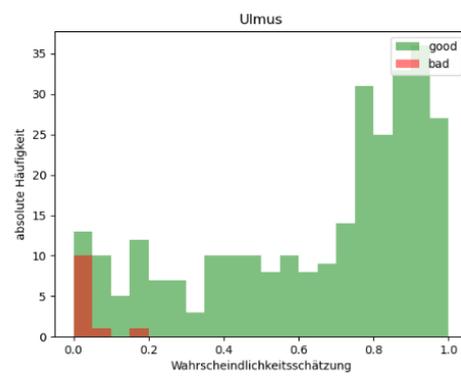
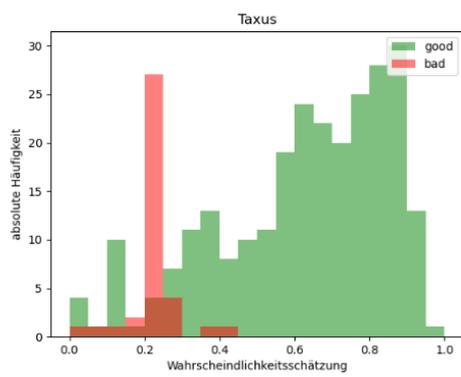
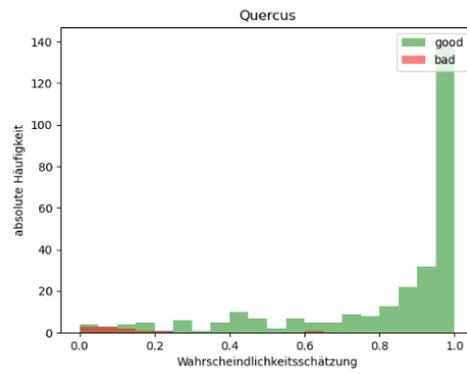
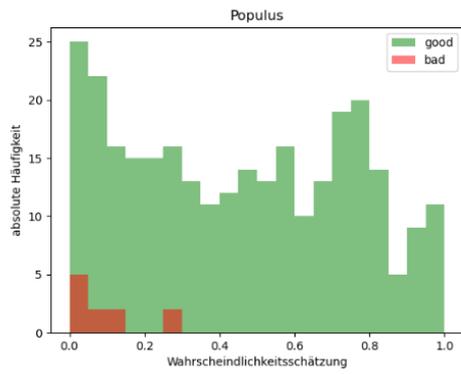
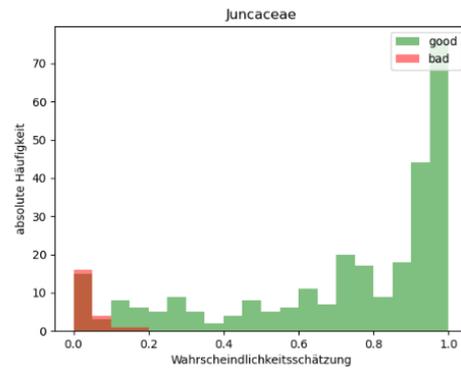
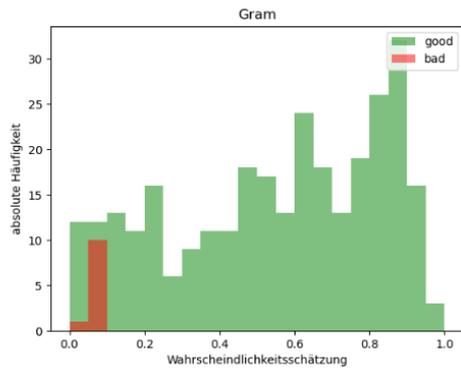
Histogramme der Wahrscheinlichkeitsschätzung Trainingsdatenset



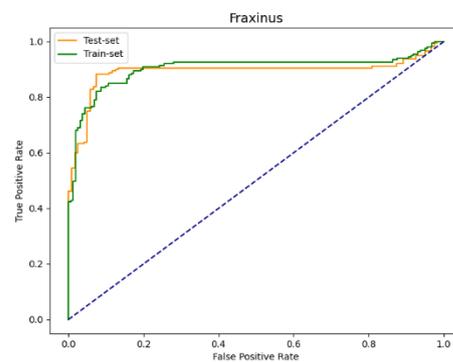
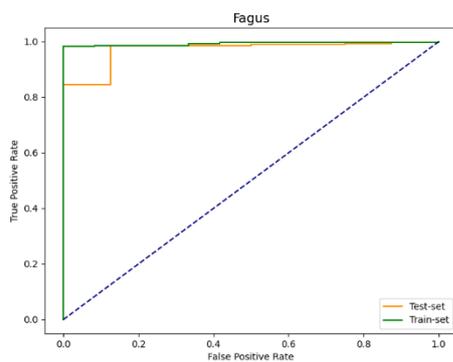
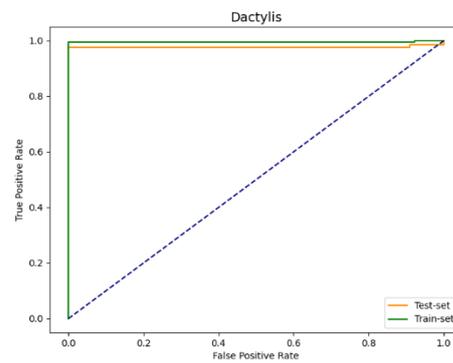
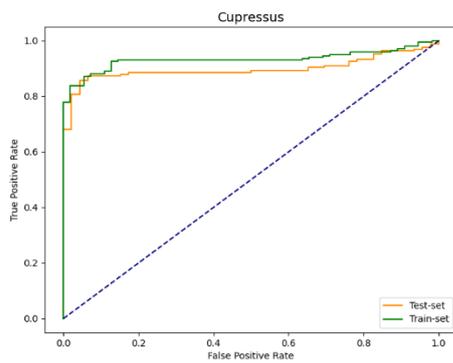
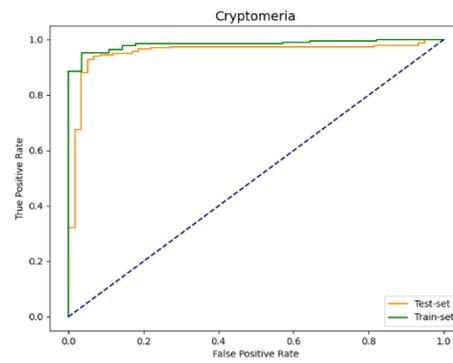
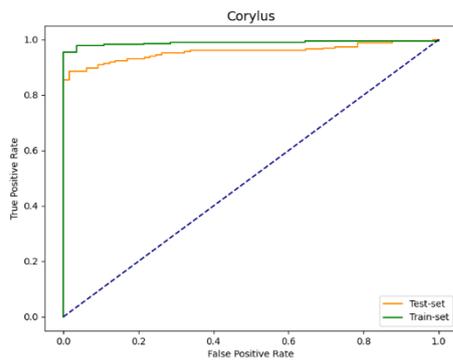
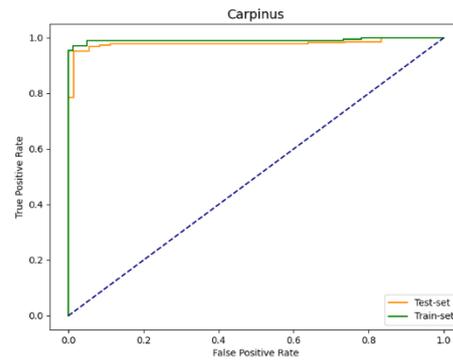
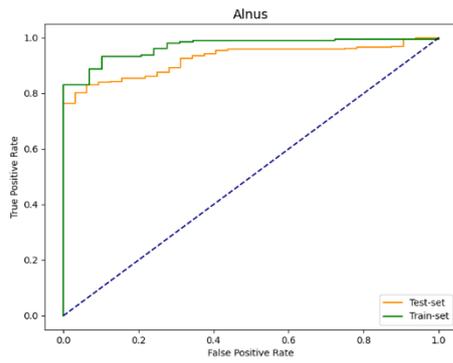


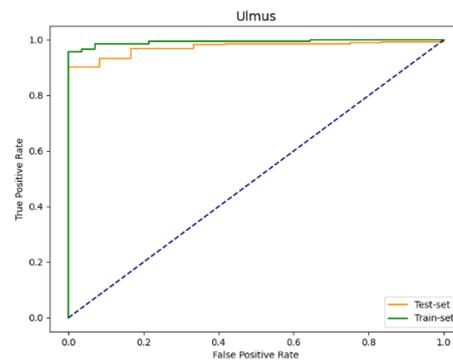
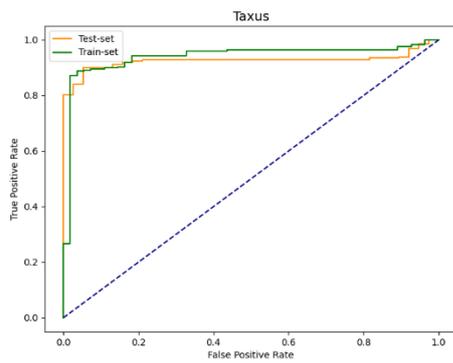
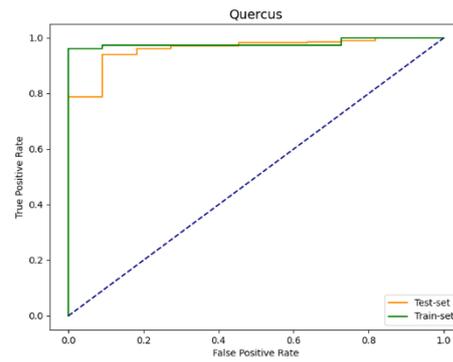
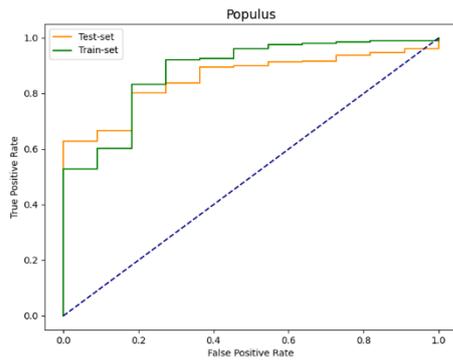
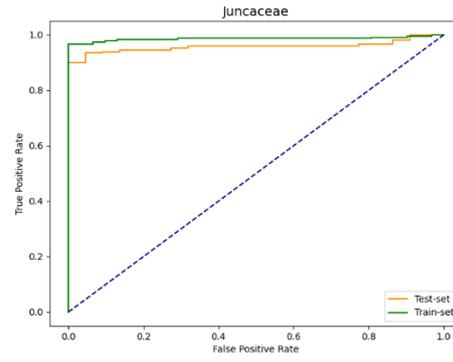
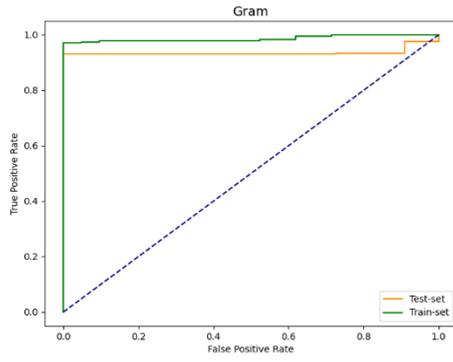
Histogramme der Wahrscheinlichkeitsschätzung Testdatenset





ROC-Kurven





Elektronischer Anhang

- A) Aufgabenstellung
- B) Projektplanung
- C) Pollenrecherche
 - 1) Merkmalliste «neu»
 - 2) Referenzlisten
 - 3) Illustrated Pollen Terms
 - 4) Quick Reference Glossary with Illustrations
 - 5) PalDat Worksheet
- D) Erstellte Datensätze
 - 1) Trainingsdatensatz gut
 - 2) Trainingsdatensatz schlecht
 - 3) Testdatensatz gut
 - 4) Testdatensatz schlecht
- E) Quellcode
 - 1) Featureextraktion, Vorverarbeitung
 - 2) PCA, Klassifikation, Aussortierung
 - 3) Infotext
- F) Grafiken
 - 1) Boxplot
 - 2) Histogramme
 - 3) Scatterplots
 - 4) Korrelationsmaps
 - 5) Histogramme automatische Aussortierung
 - 6) ROC-Kurven
 - 7) Kontrolle-FE
- G) Zwischenpräsentation