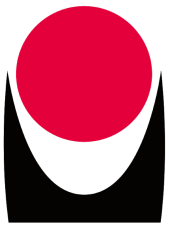


TECHNISCHE
HOCHSCHULE
LÜBECK

Intelligente Systeme

Aufgabe 2 - Identifizierung von biologischen Strukturen

Arne Bruhns & Felix Lohse

TECHNISCHE HOCHSCHULE LÜBECK

Identifizieren biologischer Strukturen

Ausgangssituation

Von der Spitze des schnabelartigen Fortsatzes am Kopf eines Segelfisches, das sogenannte Rostrum, wurden computertomographische Bilder erstellt. Anhand dieser Bilder werden Datensätze konstruiert, welche für jedes Koordinatenpaar die Höhe jenes Punktes darstellt. Mit Hilfe eines solchen Datensatzes sollen automatisiert unter Berücksichtigung von selbst analysierten Merkmalen die Mikrozähne (im Weiteren nur als Zähne bezeichnet), die sich auf dem Rostrum befinden, ausfindig gemacht und gekennzeichnet werden.

Annahmen / Limitierungen

In den Datensätzen kann es vorkommen, dass für manche Koordinatenpaare keine Höhenwerte eingetragen sind (vgl. NaN). Wir setzen diese Werte auf die Höhe Null. Der Grund dafür ist, dass wir annehmen, dass man auf den Bilder bzw. auf den Scans eh keine Zähne ausfindig machen kann, da ansonsten Höhenwerte von mindestens Null eingetragen wären. Wir können keine genauen Aussagen darüber machen, weshalb es für manche Koordinatenpaare keine Höhenwerte gibt. Eine von unserer Seite aus, aber nicht belegte Vermutung wäre, dass der Teil des Rostrum nicht gerade abgeschnitten worden ist, sondern minimal schief. Das würde bedeuten, wenn man das Rostrum längst aufschneiden, ausrollen und auf eine Platte o.ä. legen würde, dass die daraus resultierende Fläche nicht einem Rechteck entsprechen würde. Da aber die gegebenen Datensätze aus einer zweidimensionalen Matrix besteht, muss über die Fläche des Rostrum ein Rechteck gelegt werden, wodurch NaNs entstehen.

Wir gehen davon aus, dass die Liste mit den Koordinatoren der Zähne, die von Experten bestimmt worden sind, korrekt und vollständig ist. Außerdem erlauben wir bei Gegenprüfung der gefundenen Zähne mit Hilfe der Labels eine Umkreissuche mit dem Radius 5, da dies erlaubt wurde und primär bei Zähnen helfen soll, die ein Plateau bilden.

Identifizierte Merkmale

Wir haben folgende Merkmale beim Untersuchen des gegebenen Datensatzes gefunden:

1. Ein Zahn bzw. dessen Spitze ist in seinem unmittelbaren Umfeld der absolute höchste Punkt.
2. Die Steigung vom Anfang des Zahnes bis hin zu seiner Spitze hat eine gewisse Steigung, die wesentlich größer ist als bei den Höhenpunkten von Unebenheiten auf dem Rostrum. Hier ist natürlich zu beachten, dass wir nicht wissen, wo ein Zahn genau beginnt - wir schätzen dazu einheitlich, wie viele Punkte wir von der Spitze aus gesehen gehen müssen.
3. Die Summe des Höhenunterschied zwischen der Spitze des Zahnes und den umliegenden Punkten ist wesentlich größer als bei den Höhenpunkten von Unebenheiten auf dem Rostrum.

Inbetriebnahme der Software

Der Software müssen zum Start drei Argumente übergeben werden:

1. Ein Datensatz, der alle Höhenwerte der Koordinaten des Scans beinhaltet. Wird im Weiteren Verlauf das Wort “Datensatz” erwähnt, so ist damit dieser Datensatz gemeint.
2. Eine Liste, die die Koordinaten der Zähne enthält. Diese wurden von Experten ermittelt.
3. Ein numerischer Werte, der aussagt, wie die der Datensatz bearbeitet werden soll:
    a. Numerischer Werte = 0: Der Datensatz wird in Trainings- und Testdaten geteilt. Es wird anschließend mit den Trainingsdaten weitergearbeitet.
    b. Numerischer Werte = 1: Der Datensatz wird in Trainings- und Testdaten geteilt. Es wird anschließend mit den Testdaten weitergearbeitet.
    c. Numerischer Werte = 2: Der Datensatz wird nicht geteilt. Es wird mit dem ganzen Datensatz weitergearbeitet.

Weitere Einstellungen muss der Benutzer nicht vornehmen. Er erhält zum Schluss der Auswertung der Daten eine visuelle Darstellung, insofern Python mit den notwendigen Erweiterungen installiert ist (matplotlib, pandas), wo sich die richtigen Zähne als schwarz umrandete grüne Kästchen und von der Software gefundenen Zähne als weiß umrandete blaue Kreise befinden, sowie die Ergebnisse des Recalls, Precision und F-Score.

Allgemeiner Ablauf der Software

Wurde die Software korrekt gestartet, so wird anhand des dritten Arguments die Daten aufgeteilt oder nicht. Anschließend werden mögliche Zähne des (Teil-) Datensatzes mittels eines Algorithmus (s. Strategie zur Zahnfindung) ermittelt.

Ist die Suche abgeschlossen, so wird anhand der Liste mit den richtigen Zähnen unter Berücksichtigung, ob der Datensatz aufgeteilt worden ist, der Recall, die Precision und der F-Score berechnet.

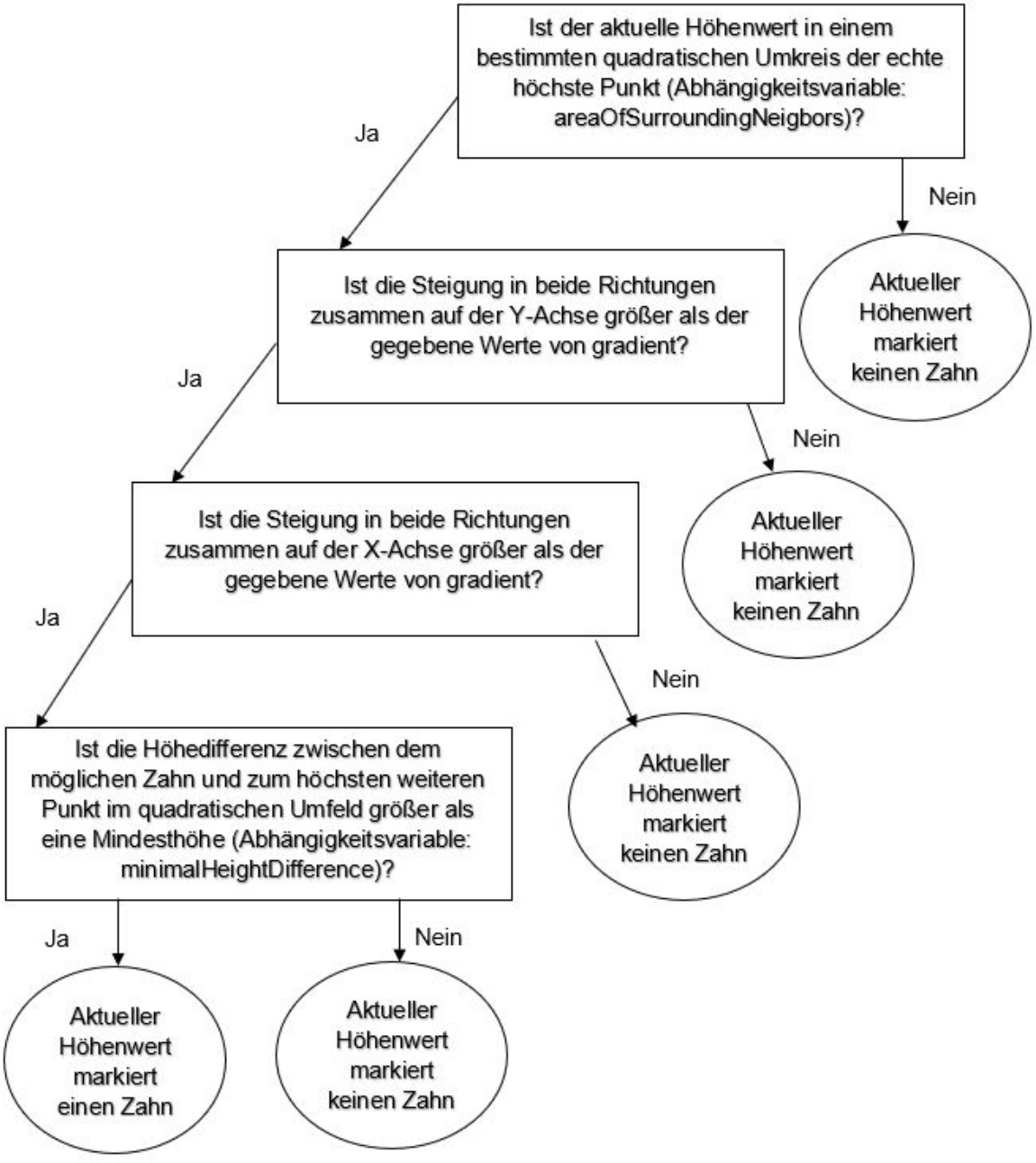
Anschließend wird eine visuelle Darstellung der Ergebnisse - wie schon im vorherigen Abschnitt erwähnt - angezeigt. Dies geschieht mittels Python. Es wird eine Heatmap erstellt, die die Höhenwerte aus dem Scans des Rostrums beinhaltet und die richtigen sowie die ermittelten Zähne markiert. Zudem wird diese Heatmap als Bild gespeichert sowie auch eine CSV-Datei mit den Koordinaten der gefunden Zähnen.

Unterteilung der Daten in Trainings- und Testdaten

Der Datensatz wird quadratisch aufgeteilt, d.h. es werden Kacheln über das zweidimensionale Array gelegt. Jedes Koordinatenpaar gehört genau zu einer Kachel. Durch einen Zufallsgenerator werden mit Hilfe eines gegebenen Seeds ungefähr die Hälfte der Kacheln ausgewählt und als Trainingsdaten gemerkt. Die verbleibenden anderen Kacheln sind automatisch die Testdaten.

Strategie zur Zahnfindung

Unsere Strategie zur Zahnfindung kann man als einen Entscheidungsbaum darstellen:



(Abbildung 1)

Damit ein Höhenwert als ein Zahn identifiziert werden kann, müssen alle Merkmale, die im Abschnitt “Identifizierte Merkmale” beschrieben sind, erfüllt sein. Wie man in Abbildung 1 sehen kann, wird ein Merkmal nach dem anderen überprüft und, sofern ein Merkmal nicht erfüllt ist, zurückgegeben, dass es kein Zahn sein kann. Ansonsten handelt es sich um einen Zahn.

Wird nur ein Teil des Datensatzes betrachtet, so wird dies in dieser Strategie berücksichtigt. Es wird bei jedem Höhenwert überprüft, ob er Teil von einer der Kacheln ist, die überprüft werden. Für die Überprüfung der Merkmale wird sich allerdings wieder auf den ganzen Datensatz bezogen. D.h. benachbarte Datenpunkte müssen nicht Teil der gleichen Kachel sein.

Parametereinstellungen

In unserer Software gibt es ein paar Parameter (u.a. auch Merkmalsparameter zur Zahnfindung), welche schnell und leicht geändert werden können. Anhand der schnellen Umstellung der Parameter kann der User schnell herausfinden, mit welchen Einstellungen die Zahnfindung erfolgreich ist.
Folgende Parameter haben wir erstellt:

1. notANumberReplacement: Wie schon im Abschnitt “Annahmen / Limitierungen” erläutert, gibt es im gegebenen Datensatz für manche Koordinaten keine Werte. Mit diesem Parameter kann man all diese Werte auf einen gewünschten Wert setzen.
2. tileSize: Mit diesem Parameter wird eingestellt, wie groß die Kacheln bei der Datenaufteilung sein sollen.
3. gradient (kurz grad): Mit diesem Parameter wird die minimale Steigung (bezogen auf ein bestimmtes Intervall von einem identifizierten Hochpunkt), welcher nicht unterschritten werden darf, damit dieser Hochpunkt als ein Zahn angesehen wird.
4. areaOfSurroundingNeighbors (kurz area): Mit diesem Parameter wird von einem Hochpunkt eine rechteckige Fläche bestimmt, wo alle Punkte dieser Fläche zu seinem unmittelbaren Umfeld gehört. Der Hochpunkt ist dabei normalerweise der Mittelpunkt der Fläche.
5. minimalHeightDifference (kurz height): Mit diesem Parameter wird eine minimale Höhendifferenz eingestellt, die sich von auf den höchsten Nachbarn des gefundenen Hochpunktes bezieht. Die minimale Höhendifferenz zwischen diesen beiden Punkten darf nicht unterschritten werden, damit der Hochpunkt als ein Zahn angesehen wird.
6. seed: Mit diesem Parameter wird der Seed für den Zufallsgenerator (siehe Unterteilung der Daten in Traings- und Testdaten) festgelegt.
7. toleranceRange: Dieser Parameter stellt einen Toleranzbereich dar. Im Klartext bedeutet das, dass ein gefundener möglicher Zahn von irgendeinem richtigen Zahn nur soweit entfernt sein darf wie der Wert des Parameters.

Zur Untersuchung der Parameter und deren Korrelation zu recall, precision und F-Score wurde eine veränderte Form der Software ausgeführt, die es ermöglicht hat, mit 32 Threads gleichzeitig die Suche nach Zähnen mit verschiedenen Parametereinstellungen durchzuführen. Hierbei wurden für die Parameter folgende Bereiche betrachtet:

- gradient: 0.0 bis 9.6 Schrittweite: 0.1

- areaOfSurroundingNeighbors: 10 bis 30 Schrittweite: 1
- minimalHeightDifference: 5 bis 20 Schrittweite: 1

Die Ausführung der Software wurde so gestaltet, dass alle Parameterkombinationen in den genannten Bereichen durchlaufen werden. Jeder Thread führte somit für drei Gradienten-,15 minimalHeightDifference- und 20 areaOfSurroundingNeighbors-Werte die Zahnfindung durch, was eine Gesamtlast von 900 Iterationen der Zahnfindung pro Thread und insgesamt 28800 Iterationen bedeutet. Die Gesamtlaufzeit betrug ca. 42 Stunden. Die Iterationen wurden grundsätzlich auf den Trainingsdaten des zuerst erhaltenen Datensatzes durchlaufen.

F-Score

Bei dieser Untersuchung wurden auf den Trainingsdaten folgende maximale F-Scores beobachtet und als Grundeinstellungen und Empfehlung in die Software übernommen (s. Abbildung 2)

	grad	area	height	recall	precision	fscore
11061	4.3	16	5	0.966057	0.89318	0.928191
11741	4.3	16	6	0.966057	0.89318	0.928191
12418	4.3	16	7	0.966057	0.89318	0.928191

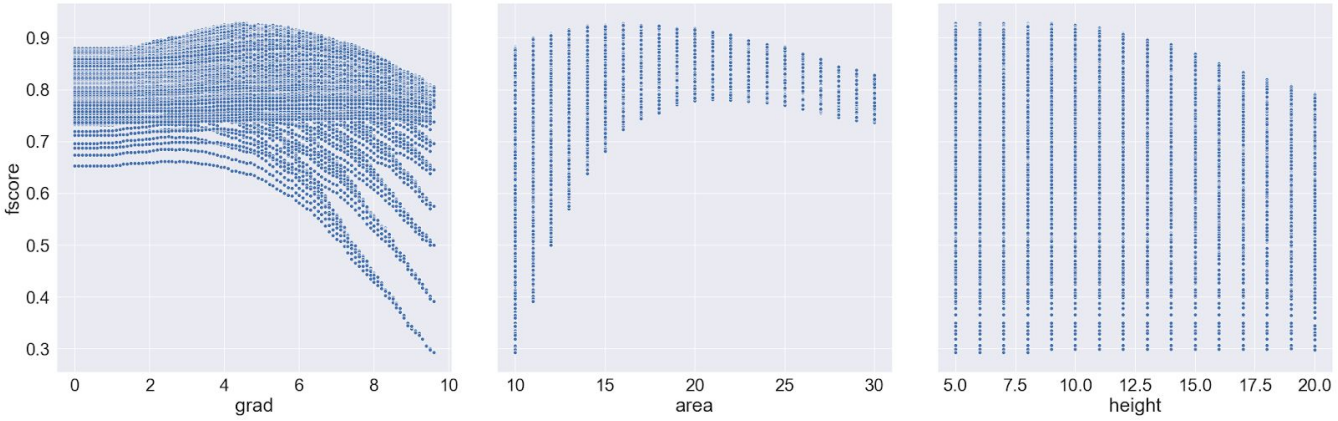
(Abbildung 2)

Aus den gesammelten Daten ließen sich die Korrelationskoeffizienten nach Pearson zum F-Score bestimmen. (s. Tabelle 1)

gradient	-0.296439
areaOfSurroundingNeighbors	0.283937
minimalHeightDifference	-0.330187

(Tabelle 1)

Abbildung 3 stellt ein Pairplot für den F-Score über die gesamten ermittelten Daten in Abhängigkeit der einzelnen Parameter dar.



(Abbildung 3)

recall

Analog zu der Darstellung der maximalen gefundenen F-Scores, trat der höchste recall-Wert bei folgenden Parameterkombinationen auf. (s. Abbildung 4)

	grad	area	height	recall	precision	fscore
11621	2.8	14	6	0.999347	0.809197	0.894276
10949	2.8	14	5	0.999347	0.808342	0.893754
12800	2.2	13	8	0.999347	0.792853	0.884204
2114	2.1	13	8	0.999347	0.785128	0.879380
23785	2.0	13	8	0.999347	0.780326	0.876359
12101	2.2	12	7	0.999347	0.777552	0.874607
12951	1.9	13	8	0.999347	0.775583	0.873360
12983	1.9	14	8	0.999347	0.774406	0.872613
2134	1.8	13	8	0.999347	0.770508	0.870134
2166	1.8	14	8	0.999347	0.769347	0.869392
710	2.1	11	6	0.999347	0.766266	0.867422
23838	1.7	14	8	0.999347	0.765117	0.866686
12991	1.6	14	8	0.999347	0.762071	0.864727
2164	1.5	14	8	0.999347	0.758296	0.862292
23461	1.4	14	8	0.999347	0.756797	0.861322
40	2.1	11	5	0.999347	0.755676	0.860596
12805	1.3	14	8	0.999347	0.755303	0.860354
2135	1.2	14	8	0.999347	0.754559	0.859871
23529	1.1	14	8	0.999347	0.753816	0.859388
12840	1.0	14	8	0.999347	0.752334	0.858424
2145	0.9	14	8	0.999347	0.750490	0.857223
23830	0.8	14	8	0.999347	0.749388	0.856503
12990	0.7	14	8	0.999347	0.749022	0.856264
2163	0.6	14	8	0.999347	0.748289	0.855785
12813	0.1	14	8	0.999347	0.747924	0.855546
2169	0.3	14	8	0.999347	0.747924	0.855546
13003	0.4	14	8	0.999347	0.747924	0.855546
23470	0.2	14	8	0.999347	0.747924	0.855546
2139	0.0	14	8	0.999347	0.747924	0.855546
23851	0.5	14	8	0.999347	0.747924	0.855546
12049	1.0	10	7	0.999347	0.712093	0.831613

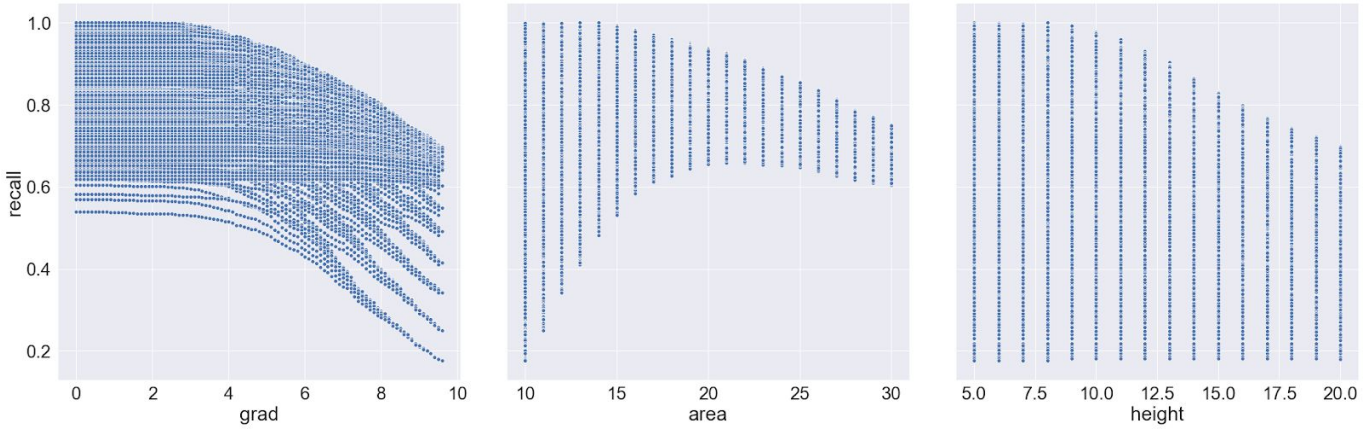
(Abbildung 4)

Hier ergaben sich ebenfalls Korrelationskoeffizienten nach Pearson zum recall. (s. Tabelle 2)

gradient	-0.476416
areaOfSurroundingNeighbors	0.139452
minimalHeightDifference	-0.418451

(Tabelle 2)

Abbildung 5 stellt ein Pairplot für den recall über die gesamten ermittelten Daten in Abhängigkeit der einzelnen Parameter dar.



(Abbildung 5)

precision

Abbildung 6 ist die Parameterkombination zu entnehmen, die die höchste precision bei der Untersuchung bot.

	grad	area	height	recall	precision	fscore
6380	9.3	20	14	0.707572	0.9627	0.815651

(Abbildung 6)

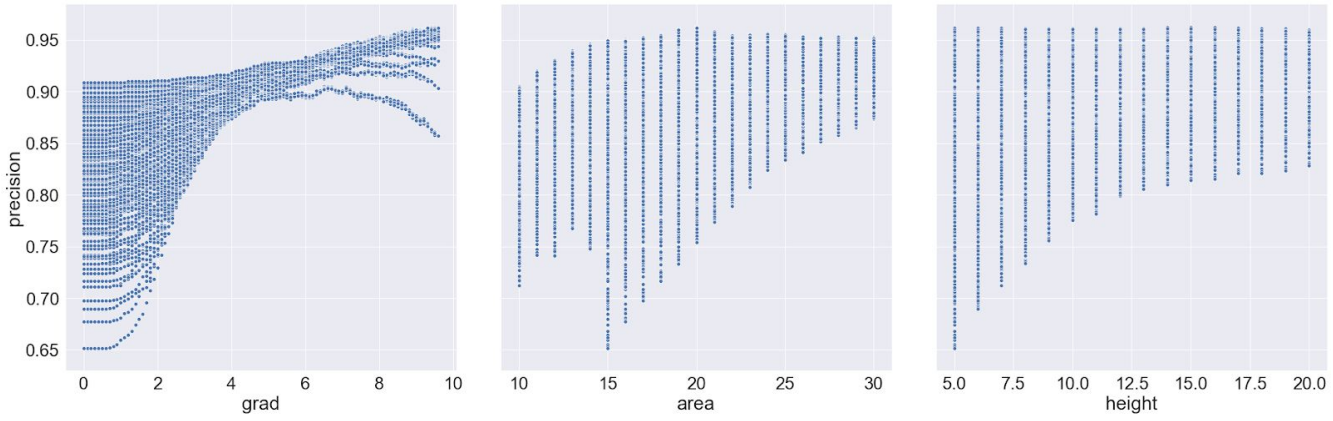
Korrelationskoeffizienten nach Pearson zur precision. (s. Tabelle 3)

gradient	0.813863
areaOfSurroundingNeighbors	0.265705
minimalHeightDifference	0.164879

(Tabelle 3)

In Tabelle 3 fällt besonders der Koeffizient für den Parameter gradient auf, der eine starke Korrelation zeigt und somit ein treffendes Merkmal darstellt.

Abbildung 7 stellt ein Pairplot für die precision über die gesamten ermittelten Daten in Abhängigkeit der einzelnen Parameter dar.



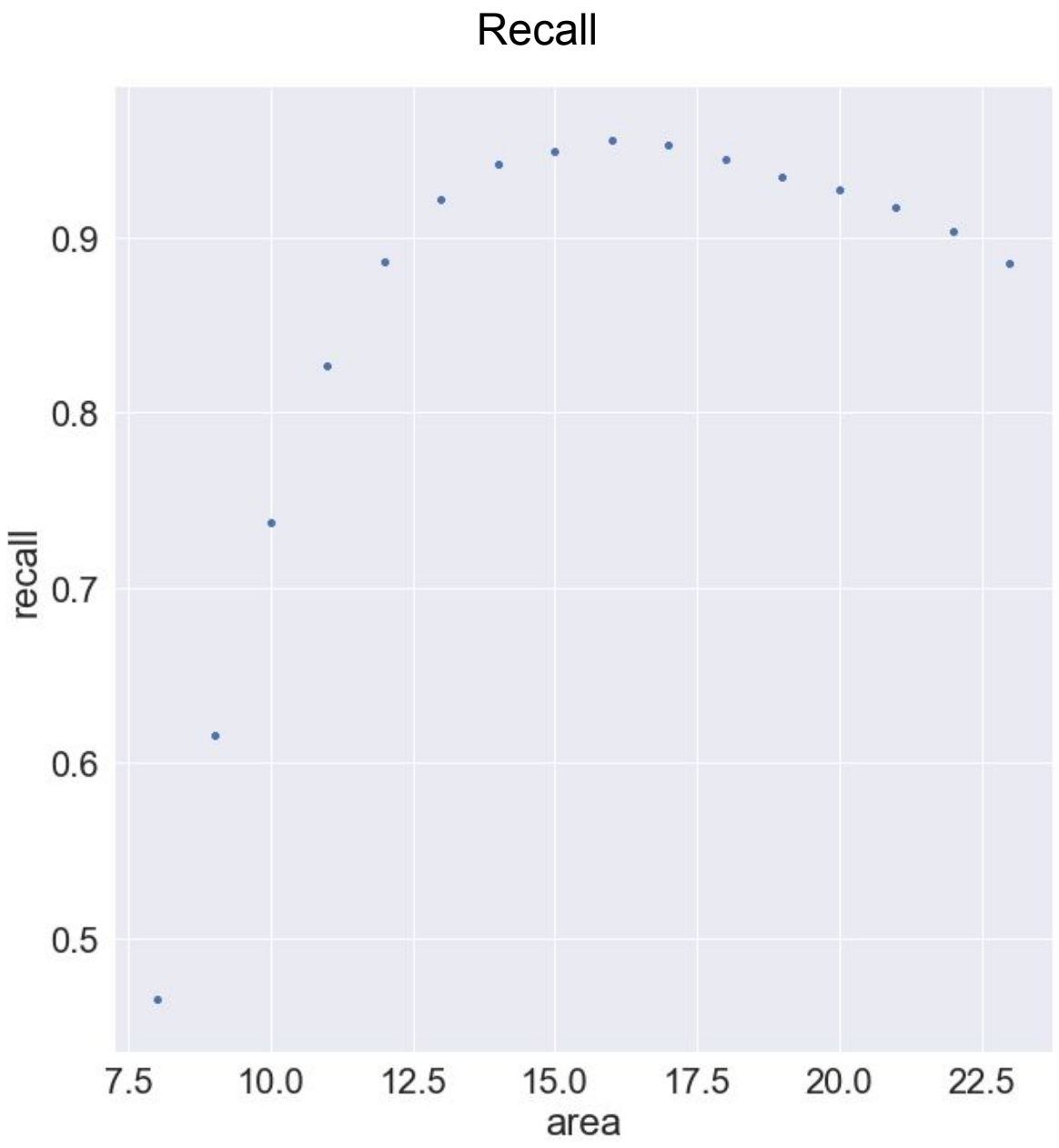
(Abbildung 7)

Bewertung

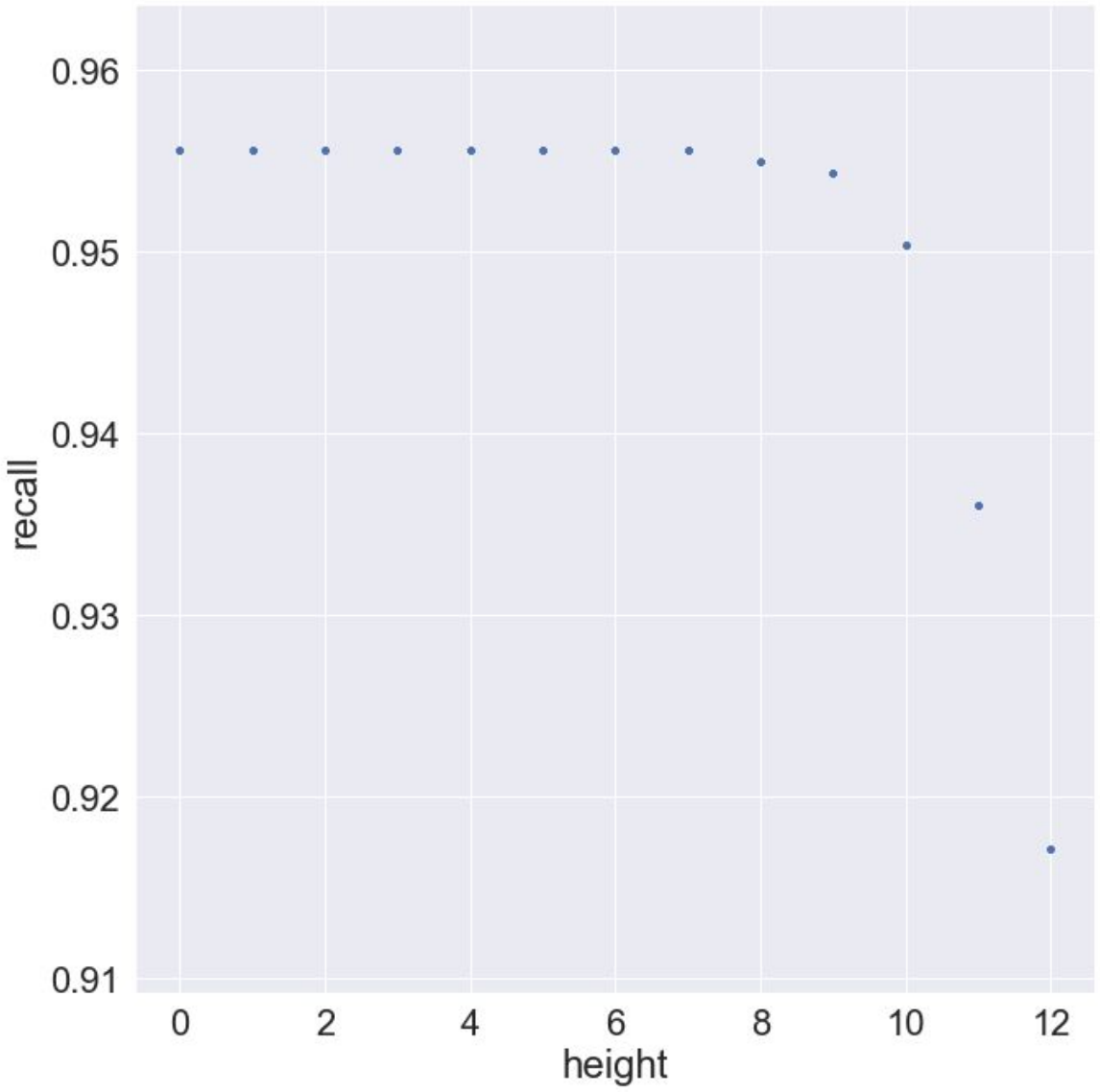
In diesem Projekt wurde kein eindeutiger, falls existent, Parameter gefunden, der eine Identifizierung erleichtert, sondern die Kombination der Parameter ist für die erfolgreiche Identifizierung notwendig.

Abweichungen von Standardeinstellungen

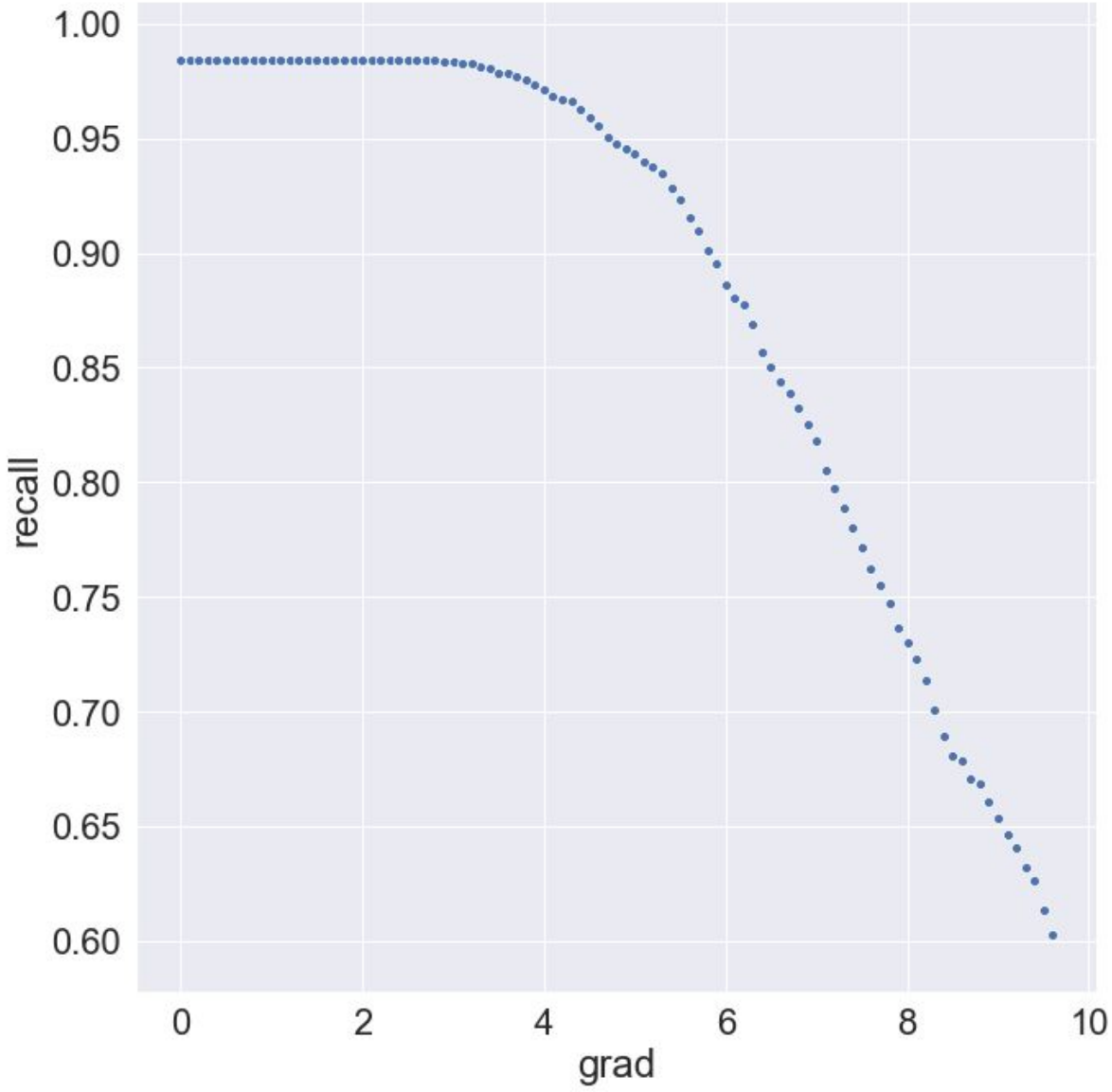
Um die Zusammenhänge der Ergebnisveränderungen zu zeigen, wenn ein einzelner Parameter von den Voreinstellungen geändert wird, sind in Abbildung 8 bis 16 die Auswirkungen auf entsprechende Metriken dargestellt.
Die x-Achse beschreibt den geänderten Parameter, die y-Achse die Auswirkung auf den dort aufgetragenen Wert.



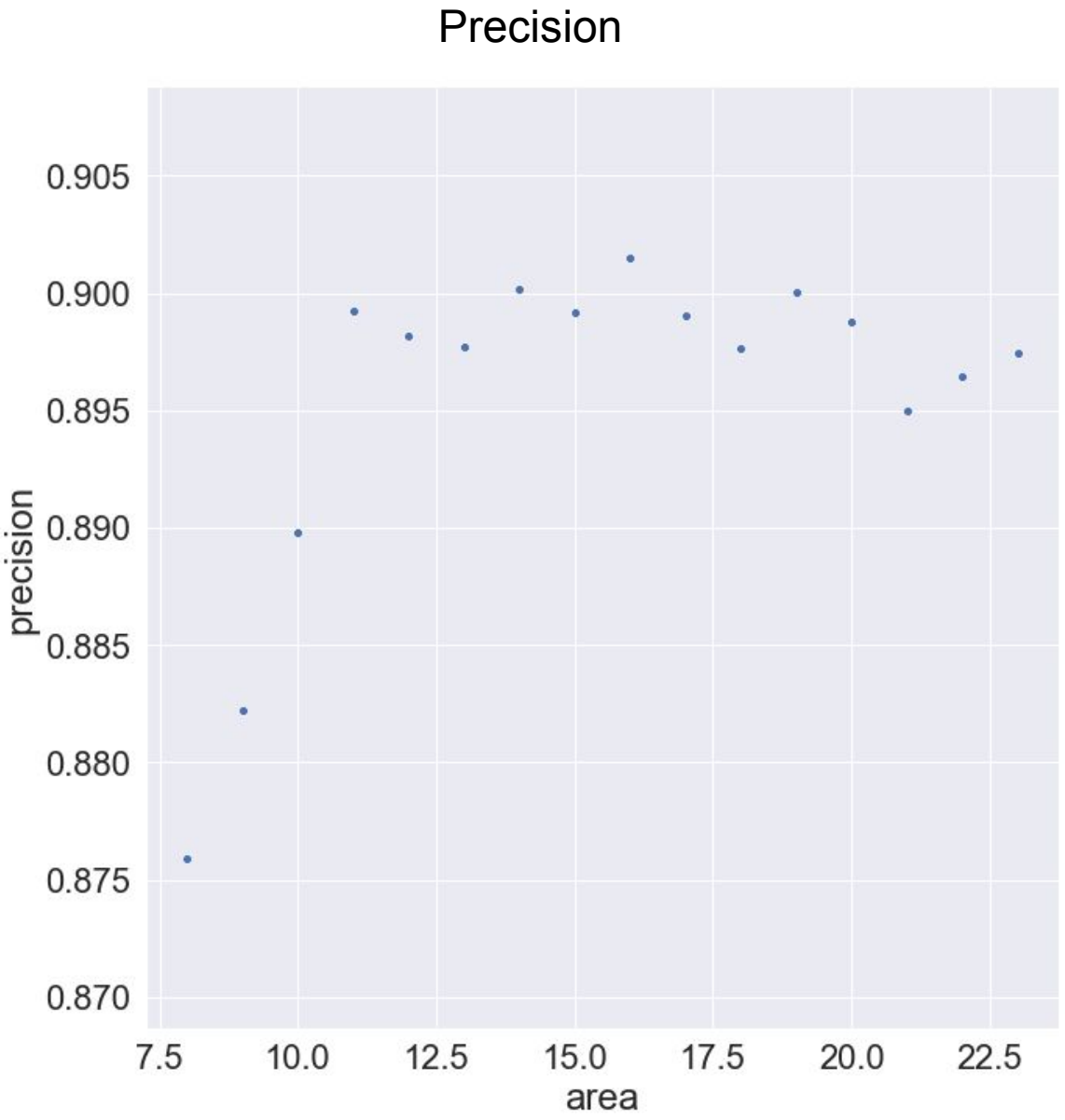
(Abbildung 8)



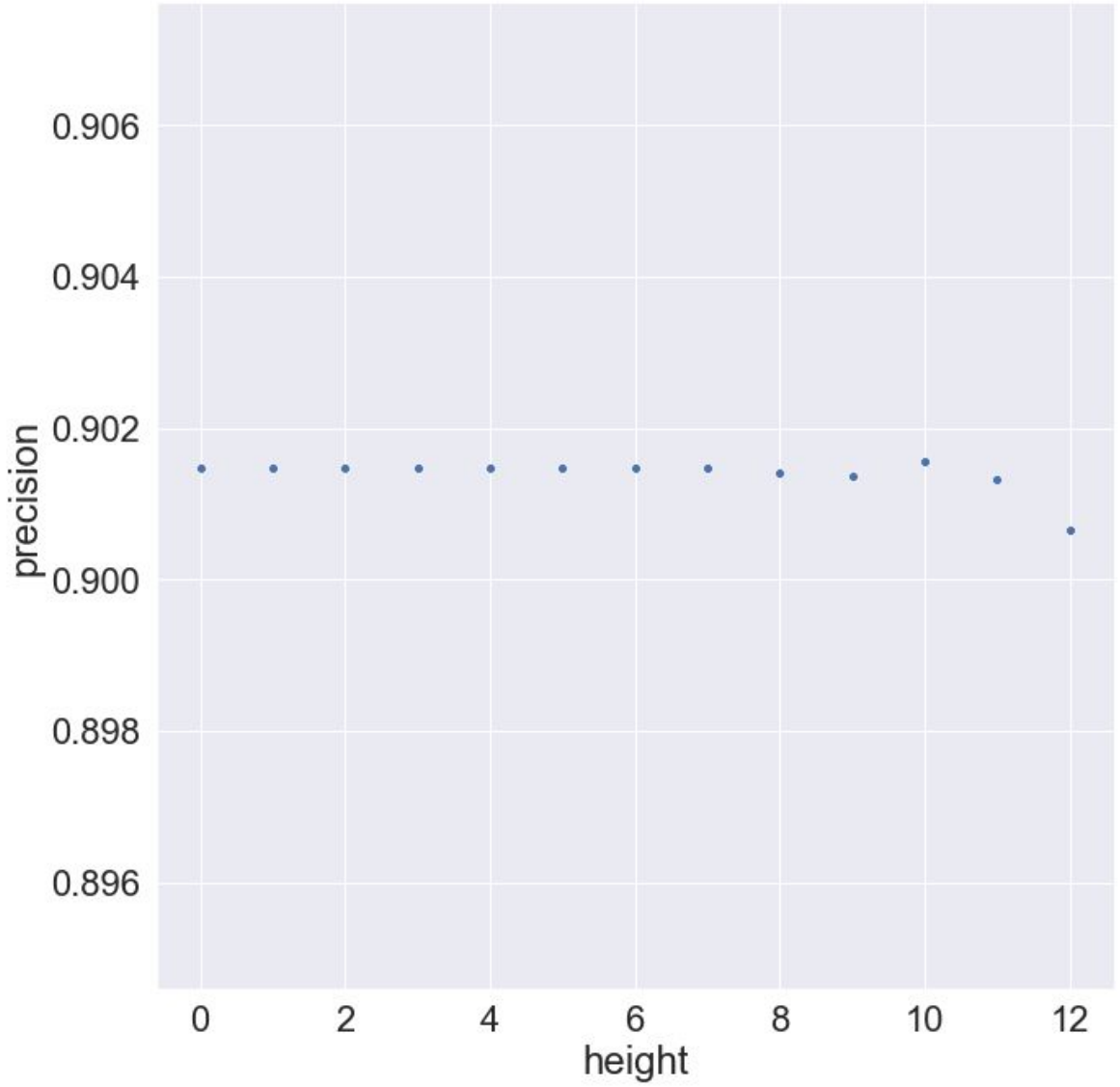
(Abbildung 9)



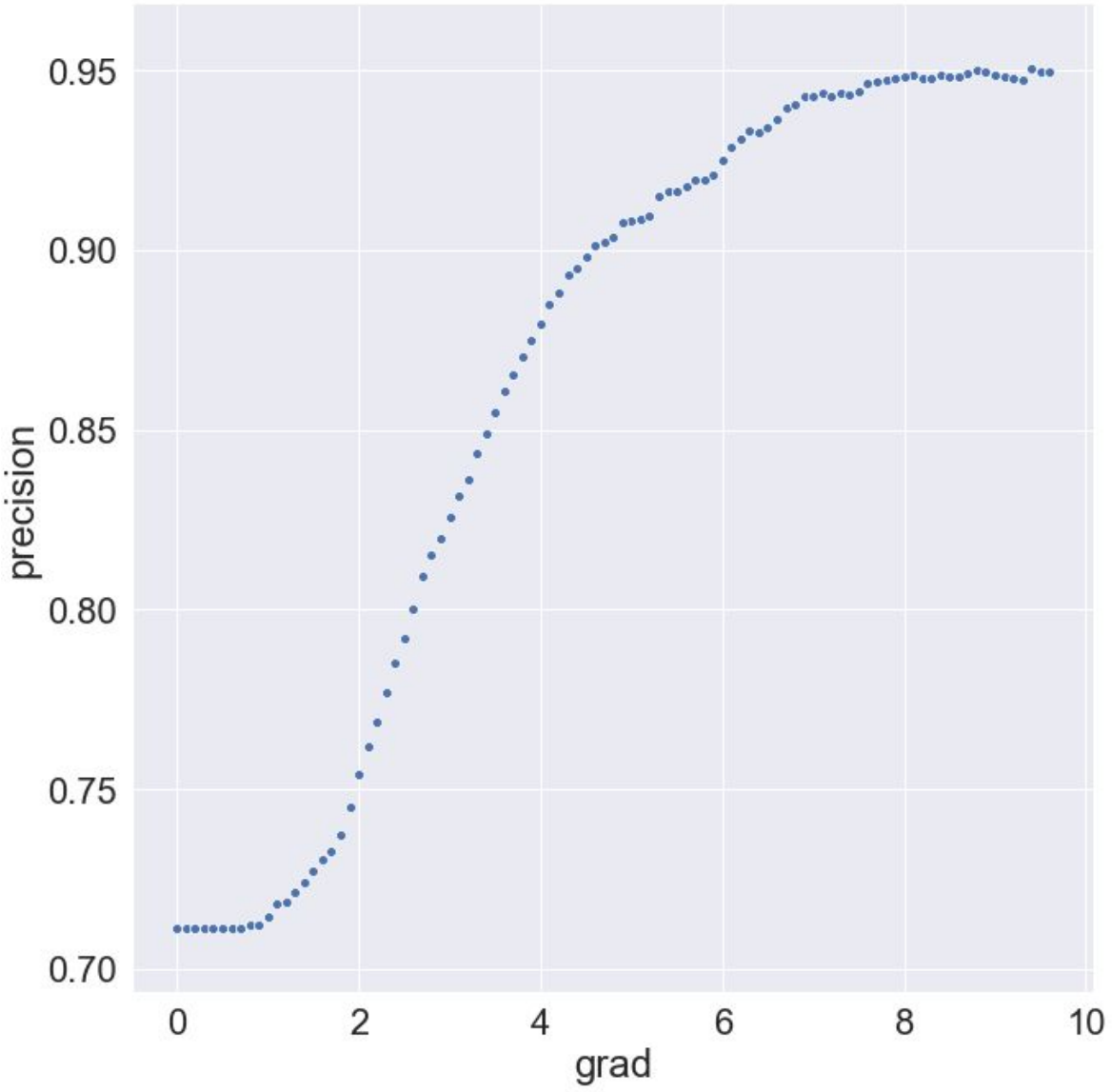
(Abbildung 10)



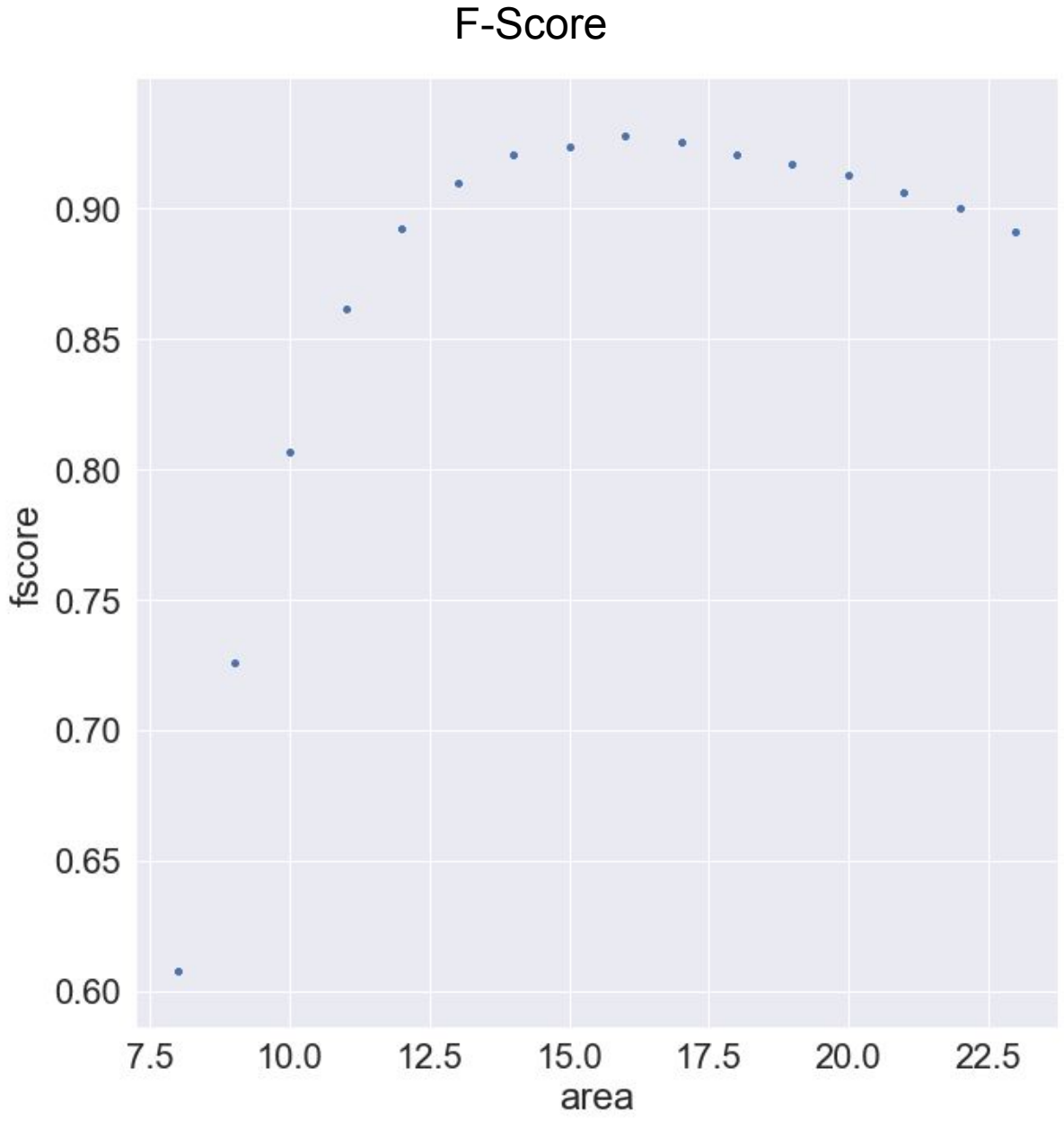
(Abbildung 11)



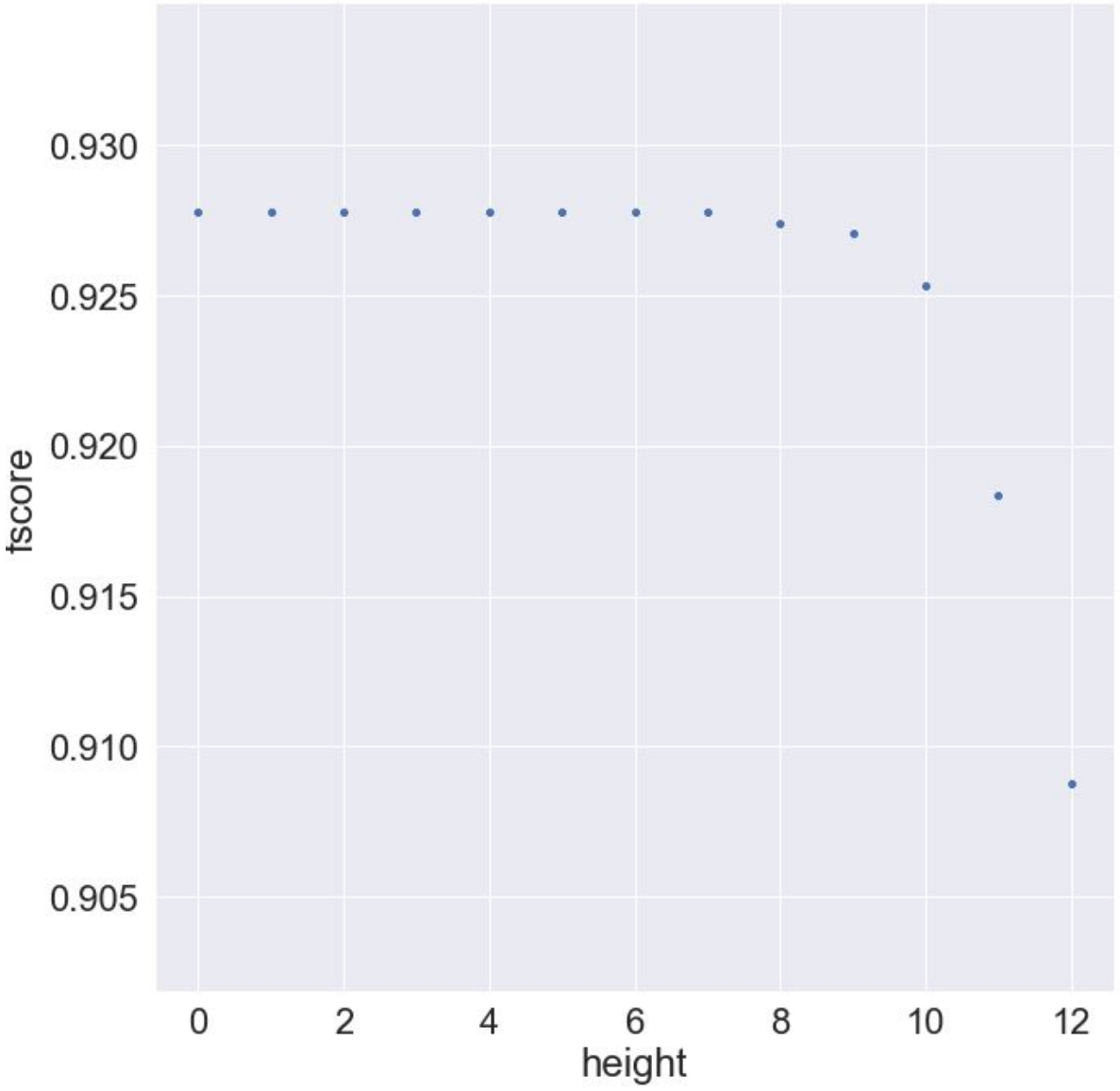
(Abbildung 12)



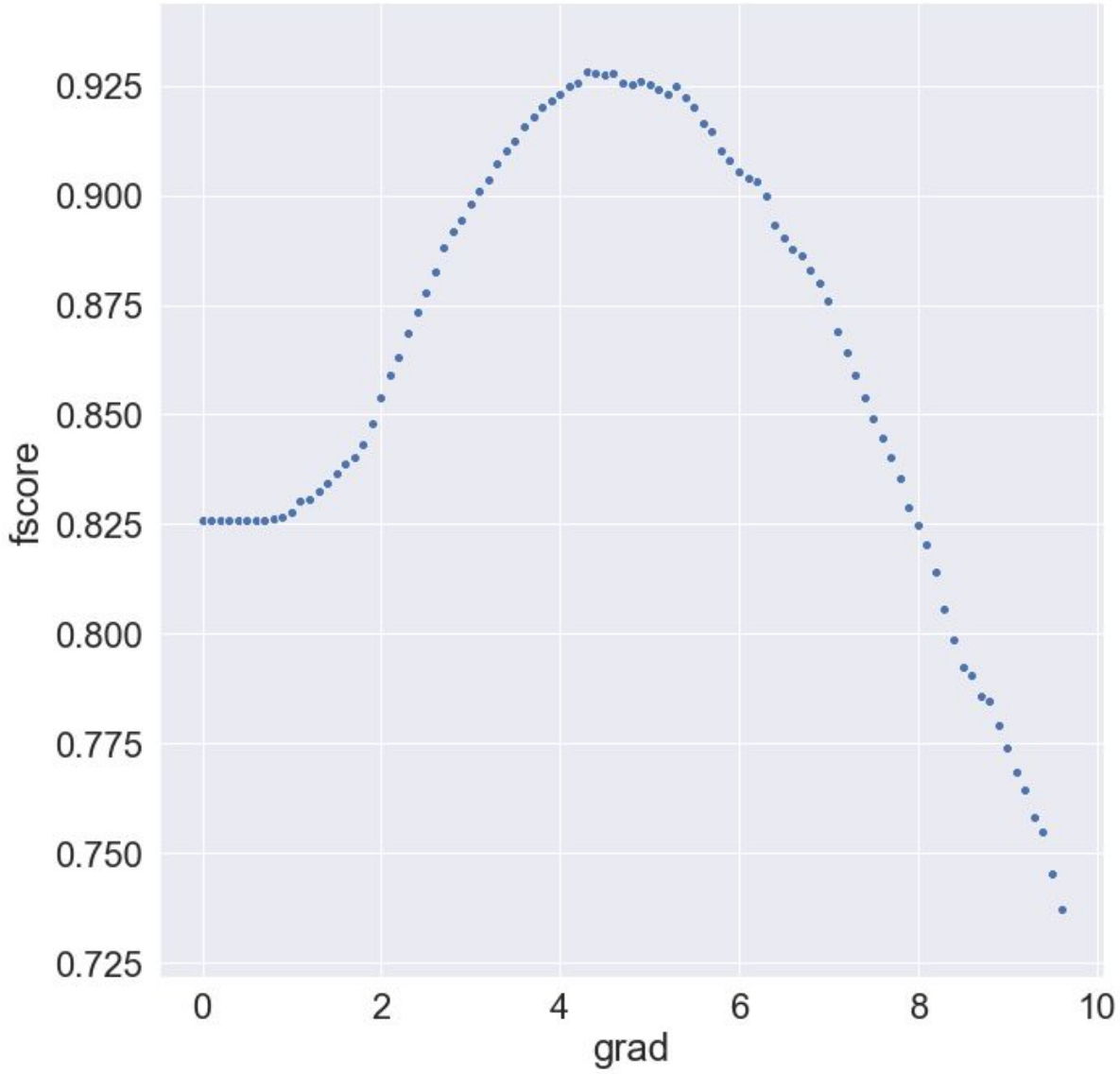
(Abbildung 13)



(Abbildung 14)



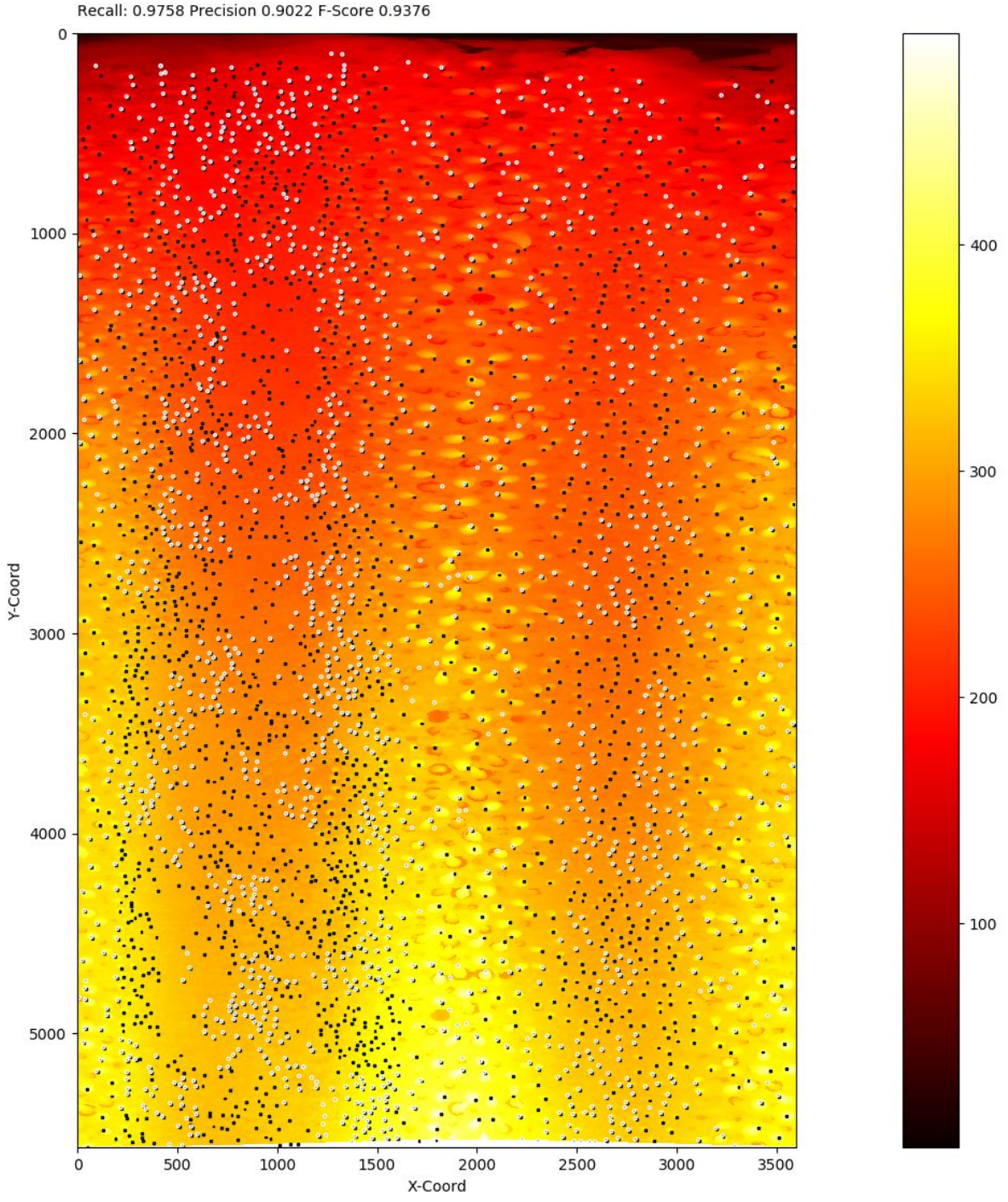
(Abbildung 15)



(Abbildung 16)

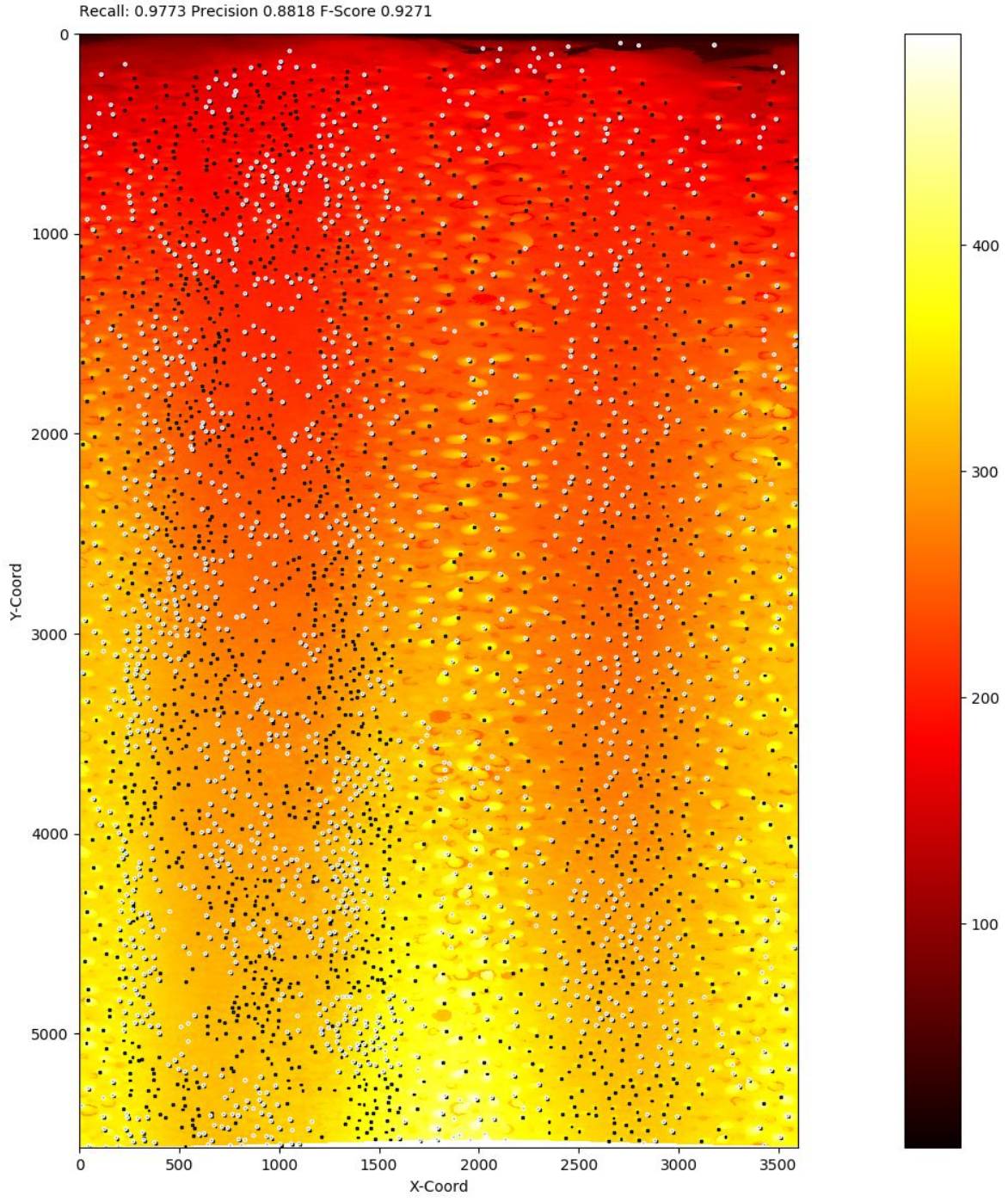
Ergebnisse und Evaluation

In Abbildung 17 ist das Ergebnis zu finden, welches nach abschließendem Training auf den Trainingsdaten ermittelt werden konnte. Besonders hervorzuheben ist hier der recht hohe recall, der zeigt, dass unter 2,5% der Zähne nicht gefunden wurden. Während des Projektes war es maßgeblich den recall hochzuhalten und erst im zweiten Schritt die Präzision zu erhöhen.



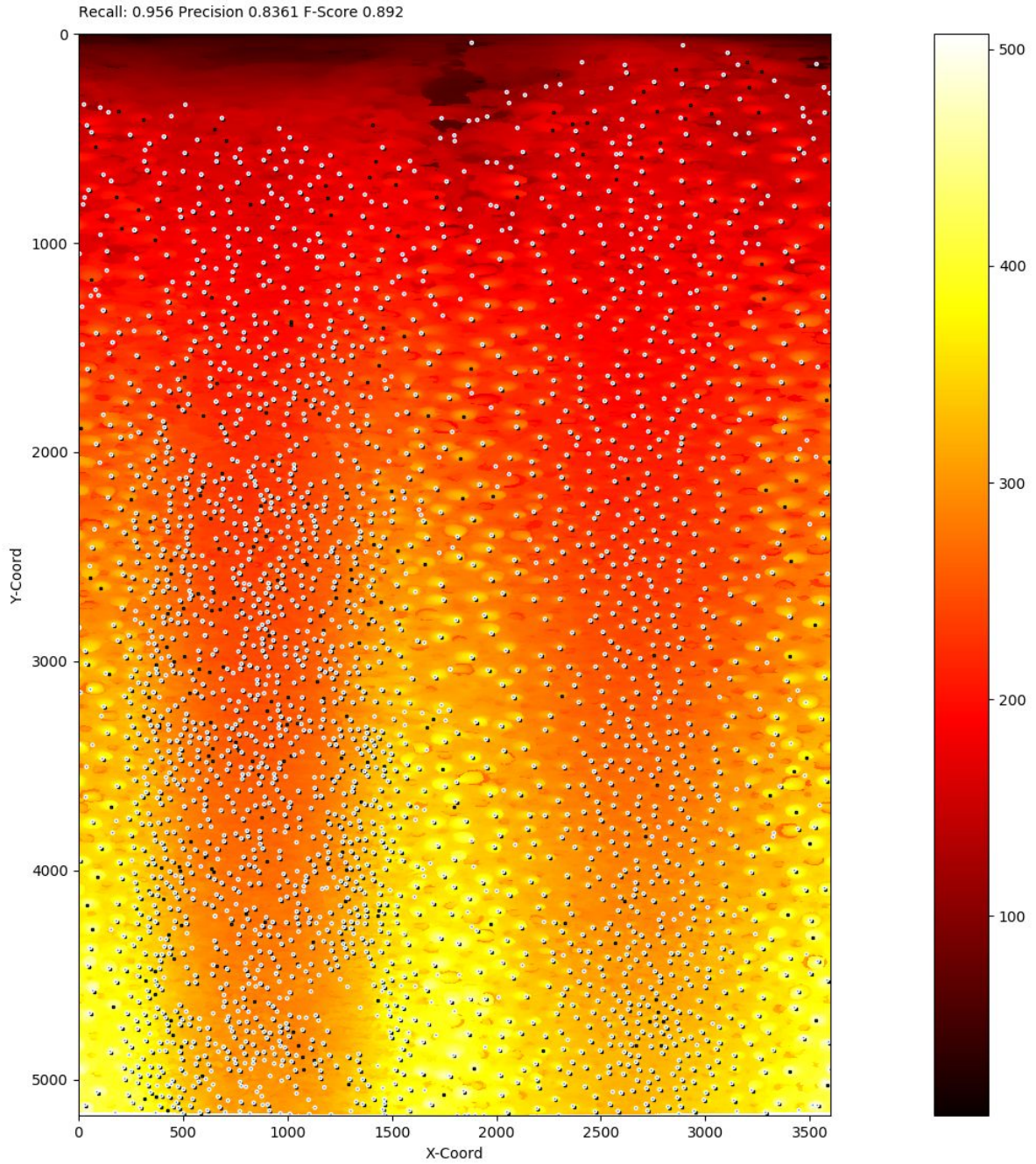
(Abbildung 17)

Abbildung 18 zeigt nun die Ausführung der Zahnfindungsstrategie auf den Testdaten des ersten Datensatzes. Trotz weiterhin hohen recalls, ist hier die Präzision unseres Algorithmus schwächer geworden, der weiterhin hohe recall, fängt diese Schwächung auf und sorgt weiterhin für einen als solide zu betrachtenden F-Score.



(Abbildung 18)

Abschließend wurde der Algorithmus auf den neu zur Verfügung gestellten Daten ausgeführt. (s. Abbildung 19) Weiterhin ist der recall relativ stabil und ist in der Ausprägung seines Wertes als gut zu bewerten. Die Präzision ist allerdings stark gesunken, was einen hohen F-Score (>0.9) verhindert.



(Abbildung 19)

Grundsätzlich arbeitet der Algorithmus vorhersehbar, es könnten noch mehr oder genauere Merkmale, vielleicht sogar das eine eindeutige Merkmal, falls existent, gesucht werden um die precision zu erhöhen. Der recall-wert zeigt, dass eine gute Basis vorhanden ist, nur die Filterung der gefundenen Zähne ist immer noch verbesserungsfähig.