VILNIAUS UNIVERSITETAS
MATEMATIKOS IR INFORMATIKOS FAKULTETAS
INFORMATIKOS INSTITUTAS
KOMPIUTERINIO IR DUOMENŲ MODELIAVIMO KATEDRA

Baigiamasis magistro darbas

Mikroplastiko dalelių aptikimas naudojant mašininio mokymosi metodus

Detection of microplastic particles using machine learning methods

Atliko:
Laurynas Čižius

Vadovas:
Dr. Tomas Raila

Vilnius
2024

Turinys

Santrauka

Šio projekto esmė yra optimizuoti metodų derinį, kuris sugebėtų efektyviai atpažinti, aptikti mikroplastiko daleles nevienodai apšviestose mikroskopijos nuotraukose. Šiame darbe taikyti binarizacijos metodai siekiantys išskirti mikroplastiko objektus nuo fono. Atlikus dalelių binarizacijos segmentaciją toliau vykdomi morfologijos sprendimai, kurie padėjo pagerinti objektų aptikimo rezultatus taip sumažinant klaidingai aptiktų dalelių atvejų skaičių. Užtikrinant tinkamą mašininio mokymosi proceso pagrindą atliekama mikroplastiko duomenų rinkinio analizė ir paruošimas, kuris skirtas mašininio mokymosi klasifikatorių modeliams. Atsižvelgiant į klasifikatorių modelių ir binarizacijos vertinimą pagal įvairius metrikos rodiklius (tikslumas, jautrumas ir F1 balansą) mašininio mokymosi sprendimas padėjo efektyviau atpažinti mikroplastiko daleles.

Summary

Detection of microplastic particles using machine learning methods

The essence of this project is to optimize a combination of methods capable of effectively recognizing and detecting microplastic particles in unevenly illuminated microscopic images. In this work, binarization methods were applied with the aim of differentiating microplastic objects from the background of the images. After the binarization process, it was observed that both the Otsu and Sauvola methods fragment microplastic particles into separate parts due to the uneven brightness and contrast in the microscopic images. Additionally, the gray spots from the filter cause an excessively high number of falsely detected microplastic particles. However, the Sauvola method is more effective because it is able to identify a larger number of true microplastic objects, despite also detecting a considerable number of falsely identified particles. After the binarization and segmentation of particles, morphological decisions were applied, which significantly improved the detection results by reducing the number of falsely detected particles.

Further refining the results obtained from the methods, a machine learning classification approach was implemented. Initially, a solid foundation for the machine learning process was established by performing an analysis and selection of features from the microplastic dataset, tailored for machine learning classifier models. Considering the evaluation of classifier models and binarization based on various metric indicators (precision, recall, and F1 score), the machine learning solution helped to more effectively recognize microplastic particles. The most optimal classifier model tested in this project is the random forest classifier, applying the hard negative mining method. This additional approach enhanced the classifier's precision in identifying true microplastic particles.

Įvadas

Plastiko taršos problema pasaulyje, iš tiesų yra viena iš didžiausių aplinkos ekosistemos problemų. Pirmojo sintetinio plastiko, kurį sukūrė Leo Baekelendas, masiška gamyba buvo pradėta 1909 metais. Per daugiau nei amžių pasaulyje susikaupė didelės plastiko šiukšlių koncentracijos , o ypač vandenynuose, upėsė, ežeruose, kurios dėl saulės spindulių, ar kitų susidėvėjimo veiksnių, suskyla į mažesnes nei 5 mm plastiko daleles, kurias dar kitaip būtų galima pavadinti - mikroplastikas [12]. Šių toksinių medžiagų kaupimasis gali būti perduodamos maisto grandine taip pasiekdamos žmogų, kadangi vandens gyvūnai (planktonai, žuvys) suvalgę mikroplastiką suserga arba miršta. Taip pat dėl taršos kinta vandens rūgštingumas, kuris griauna vandens ekosistemą [10].

Mikroplastiko taršos prevencinės priemonės yra jų vaizdų analizavimas, segmentavimas, klasifikavimas, kuris padeda formuoti veiksmingesnes strategijas ir politikos priemones, priimant informuotus sprendimus apie plastiko naudojimą ir šalinimą. Mažos plastiko dalelės fiksuojamos pagal jų formą bei dydį [33]. Tačiau ne kompiuteriniu būdų visa tai atlikti iš tiesų sudėtinga ir tai reikalauja daug laiko ir pastangų. Todėl yra ieškoma automatizuotų būdų taikyti automatinę vaizdų analizę, kad būtų aptinkami ir klasifikuojami mikroplastikai mikroskopinėse nuotraukose. Nors šioje specifinėje temoje nėra atlikta daug tyrimų, tačiau yra metodų būtent skirtų automatiniam mažų dalelių aptikimui, kategorizavimui [17]. Norint kuo tiksliau aptikti mikroplastiką svarbu turėti aukštos kokybės nuotraukų rinkinį, kadangi prastesnės raiškos, nevienodo apšvietimo mikroskopo vaizdų apdorojimas apsunkina išgauti aukšto tikslumo rezultatus.

Dalelių segmentavimas šiame projekte išgaunamas binarizavimo taikymu, tai procesas, kai originalios nuotraukos iš anksto nustatomos pagal tam tikrą slenkstinį lygį (angl. *threshold level*), pikseliai yra konvertuojami į skaitmeninį vaizdą, kuris susideda tik iš dviejų spalvų – juodos ir baltos. Pagal šias spalvas skaitmenizuotos nuotraukos skirstomos į dvi kategorijas: mikroplastikas ir vaizdo fonas. Literatūroje pabrėžiama, kad mikroplastiko dalelių išskyrimas nuo fono yra viena iš pagrindinių segmentavimo procesų [17].

Sprendžiant binarizacijos problemas mikroskopinėse nuotraukose, morfologijos vaizdų apdorojimas modifikuoja vaizdų pikselius, padėdamas pagerinti objektų išskyrimą, pašalinant ar pridedant dalelių pikselius išryškinami svarbiausi objektai, taip užtikrinant tikslesnį dalelių identifikavimą. Tai svarbu norint koreguoti faktinių objektų formų deformacijas, kurios gali kilti būtent dėl binarizavimo proceso [13].

Mašininis mokymas dalelių aptikime suteikia galimybę efektyviau klasifikuoti mikroplastiko dalelių formas, kurių rankinis identifikavimas būtų sunkus ar neįmanomas. Šis procesas nuolat tobulina klasifikavimo tikslumą mokantis iš naujų duomenų, vertinant dalelių išskirtas savybes, o naudojantis šiuos pažangius analizės modelius, aptinkamas tikslesnis mikroplastiko dalelių atvejų kiekis nei taikant vien binarizavimo sprendimą [19].

Darbo tikslas - išanalizuoti ir tobulinti mikroplastiko dalelių aptikimo bei mašininio mokymo metodus, nevienodai apšviestose optinės mikroskopijos nuotraukose.

Šiam tikslui pasiekti keliami tokie uždaviniai:

- Atlikti binarizavimo metodų tyrimus ir įvertinti jų efektyvumą mikroplastiko aptikimui.
- Pritaikyti papildomas morfologines operacijas, siekiant spręsti binarizavimo metodų trūkumus.
- Pritaikyti mašininio mokymo klasifikavimo metodus, siekiant rasti efektyviausius modelius mikroplastiko aptikimui.

1. Mikroplastikas ir jo vaizdų analizė

Mikroplastikas yra smulkios plastiko dalelės turinčios įvairias formas nuo visiškai sferinių iki ilgų pluoštų, įvairiausių spalvų, kūrių dydis būna mažesnis nei 5 mm. Pačios dalelės skirstomos pagal mikroplastiko kilmę:

- Pirminę - jau pagamintos tiesioginiam arba netiesioginiam naudojimui skirtos priemonės - dažai, kosmetika, dervos granulės, dantų pasta, tekstilės pluoštai ir begalės kitų priemonių.
- Antrinę - natūraliai susidėvėjusios, suskilusios plastiko dalys - žvejybos tinklai, maišeliai, buteliai ir kiti kasdieniai daiktai iš plastiko.

Šios dalelės yra randamos jūrose, vandenynuose, dirvožemyje ir net ore, todėl plačiai paplitusios dalelės turi būti tiriamos, kad galėtume suprasti jų potencialų poveikį žmonių sveikatai ir globaliai aplinkai [30].

Toliau siekiant įgyvendinti projekto tikslą, šiame 1. skyriaus poskyriuose detaliai nagrinėjami vaizdų paruošimo, binarizavimo, segmentavimo metodai ir jų veikimo principai.

1.1. Vaizdo paruošimas

Prieš pradedant nuotraukuose esančių objektų atpažinimą svarbu sutvarkyti vaizdų geometrinį iškraipymą, tai yra pašalinti, triukšmus vaizduose ir paruošti tinkamai objektų segmentacijai. Skaitmeninis vaizdų apdorojimas būtent tai ir leidžia išgauti ir pabrėžti svarbiausias norimas detales vaizduose, vaizdo paruošimą galime suskirstyti į 5 kategorijas [29] [17]:

- Vaizdo reprezentacija - nuotraukų skaitmeninis vaizdas susideda iš trijų spalvų: raudonos, žalios ir mėlynos (angl. *RGB*) spalvos, atspindinčias pikselio šviesumą kiekvienai pieš tai minėtos 0 - 255 spalvos diapazone. Toks platus spalvos pasirinkimas yra trukdis vaizdo apdorojimui, todėl nuotraukos yra paverčiamos pilko fono vaizdu. Be to, skaitmeniniuose vaizduose naudojamos koordinatės, kad būtų galima nurodyti konkretaus pikselio vietą vaizde.
- Vaizdo pirminis apdorojimas - pirminio apdorojimo metu galima pašalinti vaizdo triukšmą. Vaizdų triukšmas dažniausiai yra kilęs dėl tam tikrų techninių trūkumų (vaizdo kokybės stygiaus).
- Vaizdo išryškinimas - šia kategorija vaizdas yra modifikuojamas taip, kad jo kokybė po binarizavimo metodo būtų tinkama. Tai gali būti naudojama mikroplastiko dalelių vaizdo sričių kontrasto ir šviesumo korekcijai, kad vaizdas būtų aiškesnis ir detalesnis.
- Vaizdo antrinis apdorojimas - atlikus binarizavimą nuotraukoje lieka tik dveji pikseliai 0, kai nėra jokio šviesumo- juodos spalvos ir 1 kai pikselis yra baltas. Po to integruojant objektų aptikimo metodą šiame etape išskiriamos norimos vaizdo detalių formos, kurios leidžia suprasti ir interpretuoti vaizdo turinį.
- Vaizdo analizavimas - po vaizdo binarizavimo ir dalelių aptikimo svarbu įvertinti aptiktų mikroplastiko dalelių tikslumą. Dalelių tikslumo įvertinimas reikalingas tam, kad sužinotume kiek teisingai yra aptikti objektai.

1.2. Vaizdo glodinimas

Triukšmai dažniausiai atsiranda dėl nekokybiškos fotografijos, blogo apšvietimo, todėl vaizdo glodinimas mikroplastiko binarizavimui yra reikalingas pagerinant tyrimo rezultatus. Sumažinus triukšmo kiekį nuotraukoje išryškiname objektų kontūrus ir išskiriame nuo panašaus fono kontrasto spalvų ir visa tai padidina tolimesnį vaizdo binarizavimo tikslumą ir efektyvumą.

Šiame projekte pasirinktas plačiai naudojamas Gauso suliejimo (angl. *Gaussian blur*) metodas, kuris modifikuoja pradinio vaizdo pikselių reikšmes siekiant pašalinti triukšmą. Pagrindinis šio metodo veikimo principas - konvoliucija. Iš pradžių yra pasirenkamas norimas branduolio (angl. *kernel*) dydis MxN, tai yra matrica, kuri slenka per visą vaizdą. Šitaip kiekvienos matricos centrinio pikselio naujoji reikšmė apskaičiuojama konvoliucijos principu (dviejų matricų pikselių reikšmių sandaugų sumą). Didelis branduolio dydis gali lemti per didelį vaizdo suglodinimą, todėl dažniausiai pasirenkami mažesni branduoliai 3x3 arba 5x5, kurie pašalina nedidelį kiekį triukšmo ir išlaiko vaizdo ryškumą [11]. Žemiau pateiktoje 1.1 formulėje galime pamatyti Gauso suliejimo matematinę išraišką:

$$G(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}. \tag{1.1}$$

- $G(x,y)$ - Gauso funkcijos reikšmė pikselio (x,y) koordinatėse.
- σ - Nuokrypio konstanta, kuri nurodo kaip stipriai suglodinama nuotrauka.
- x ir y - Pikselio koordinatės.

1.3. Binarizavimas

Vaizdo segmentavimo pradžios metodas (binarizavimas) yra esminis žingsnis automatinėje vaizdų analizėje. Binarizavimo pagrindinis uždavinys yra atskirti objektų ir vaizdo fono elementus nuotraukose, kitaip tariant, tai yra vaizdo transformacija, iš pilkos skalės pakeičiant į dvejetainį formatą, kuriame pikseliai yra paverčiami juoda arba balta spalva. Dvejetainio reikšmė 0 arba 1 yra priklausomi nuo tam tikros slenksčio vertės, kuri yra pasirenkama remiantis globaliu arba lokaliu (adaptyviu) būdu. Globalūs (1.3.1. skyrelis) slenksčio nustatymo metodai nustato vieną slenkstį visam vaizdui, o adaptyvūs (1.3.2. skyrelis) metodai apskaičiuoja slenkstinę vertę kiekvienam vaizdo pikseliui [32]. Žemiau pavaizduotoje (1.2) formulėje galime pamatyti bendrinę proceso formulę:

$$G(x,y) = \begin{cases} 0, f(x,y) < T \\ 1, f(x,y) > T. \end{cases} \tag{1.2}$$

- $G(x,y)$ - Galutinė vaizdo išvesties pikselio reikšmė.
- $f(x,y)$ - Pradinio vaizdo pikselio reikšmė.
- T - Slenksčio vertė, automatiškai parinkta, pagal lokalų arba globalų metodą.

1.3.1. Otsu binarizavimo metodas

Šis 1979 m. globalus N. Otsu plačiausiai naudojamų vaizdų binarizavimo technikų suranda vieną bendrą slenkstinės ribos reikšmę visiems nuotraukos pikseliams [24]. Ši reikšmė nustatoma

pasitelkiant vaizdo pikselių histogramą, tai suskirsto vaizdo pikselių spalvų tonus ir apskaičiuoja, kiek kiekviename intervalo 0 - 255 ribose (8 bitų gylis, pilko fono nuotrauka) yra būtent to atspalvio pikselių. Tuomet Otsu metodas sudaręs dvi pikselio klases- fonas ir objektai, skaičiuoja kiekvienos įmanomos slenkstinės reikšmės vidinės klasės dispersiją, siekiant atrasti jos minimaliausią reikšmę tarp klasių. Galiausiai, visi pikseliai žemesni už pasiektą optimaliausią slenkstinę reikšmę priskiriami fonui, o aukštesni objektams.

1.3.2. Sauvola binarizavimo metodas

Adaptyvusis Sauvola binarizavimo metodas randantis kiekvieno vaizdo pikselio optimaliausią slenkstinę vertę buvo publikuotas moksliniame darbe 2000 m. autorių J. Sauvola ir M. Pietikäinen [31]. Pats metodo veikimo principas pagrįstas slenkamuoju langu, kuris slenkamas per kiekvieną vaizdo pikselį. Šis slenkantis langas kiekvienam pikseliui apskaičiuoja šviesumo intensyvumo vidurkį, kuris parodo būtent konkretaus pikselio vaizdo srities šviesumą, kontrastą tarp objektų ir fono. Pats lango dydis pasirenkamas atsižvelgiant į konkrečią užduotį (projektinėje dalyje bus nuspręsta koks lango dydis optimaliausias mikroplastiko dalelių atpažinimui atlikti). Kartu su intensyvumo vidurkiu lange apskaičiuojamas standartinis nuokrypis, kuris parodo pikselių šviesumų pasiskirstymo įvairovę. Šis apskaičiuotas nuokrypis padeda įvertinti, kaip stipriai atskiri pikseliai nukrypsta nuo vidutinio šviesos kontrasto intensyvumo, slenkančiame lange. Prie šių dviejų nefiksuotų parametrų dar yra naudojami Sauvola slenkstinės vertės formulės (1.3) R ir k parametrai, kur R nustato didžiausią galimą standartinį nuokrypį, o k- konstanta reguliuoja binarizavimo jautrumą. Dėl mažesnės k vertės binarizavimas tampa mažiau jautrus vaizdo triukšmui ir pikselio šviesumo, spalvos intensyvumo svyravimams, todėl vaizdo objektų (mikroplastiko dalelės) struktūros gali būti binarizavimo metu priskirtos prie fono. Visi šie paminėti procesai žemiau pateiktoje formulėje (1.3) apskaičiuoja kiekvieno lango centrinio pikselio slenkstinę vertę:

$$T(x,y) = m(x,y)\left(1 + k\left(\frac{s(x,y)}{R} - 1\right)\right). \tag{1.3}$$

- $T(x,y)$ - Vieno pikselio gauta slenkstinė vertė.
- $m(x,y)$ - Intensyvumo vidurkis slenkamajame lange, esančiame apie centrinį pikselį (x,y).
- $s(x,y)$ - Standartinis nuokrypis slenkamajame lange, esančiame apie centrinį pikselį (x,y).
- R - Maksimali standartinio nuokrypio reikšmė.
- k - Konstanta, reguliuojanti slenkstinio intensyvumo vertę.

1.4. Objektų kontūrų išskyrimas

Po binarizavimo proceso, toliau siekiant automatizuotai identifikuoti mikroplastiko daleles, būtina aptikti ir išskirti objektų briaunas, kontūrus. Visa tai suteiks informaciją apie objekto formą, dydį ir kitas geometrines savybes, kurios yra svarbios mikroplastiko dalelių atpažinimo ir klasifikavimo užduotyse. Egzistuoja daugiau nei vienas briaunų radimo metodų [26], vienas iš optimaliausių - Canny metodas. Šis John F. Canny 1986 m. sukurtas algoritmas integruoja vaizdo glodinimo, gradientų intensyvumo radimą ir kraštų sekimo etapus [5]. Canny vaizdo briaunų aptikimo proceso pradžioje yra įgyvendinamas vaizdo glodinimas naudojant Gauso filtrą (šis metodas

apžvelgtas 1.2. skyriuje), tai integruotas filtravimas esančiuose Canny programavimo bibliotekose, tačiau dvigubas filtravimas gali būti naudingas esant dideliam triukšmo lygiui, nes tai gali padėti dar labiau sumažinti triukšmą, tinkamiau išskiriant objektų briaunas. Sekančiame etape skaičiuojamas gradientas kiekvienam vaizdo pikseliui naudojant horizontalius ir vertikalius konvoliucijos Sobel operatoriaus 3x3 dydžio filtrus (konvoliucijos skaičiavimas vyksta lango principu slenkant per vaizdo pikselius). Tai leidžia nustatyti kiekvieno pikselio intensyvumo gradiento stiprumą (formulė 1.4) ir kryptį (formulė 1.5) vaizde, kurie identifikuoja briaunų vietą ir jų krypties orientaciją.

$$G = \sqrt{G_x^2 + G_y^2}. \tag{1.4}$$

$$\theta = \arctan\left(\frac{G_y}{G_x}\right). \tag{1.5}$$

- G - Galutinis pikselio gradiento stiprumas.
- θ - Gradiento kryptis.
- G_x - Vaizdo pikselių ir horizontalaus Sobel operatoriaus konvoliucijos būdu apskaičiuotas gradientas.
- G_y - Vaizdo pikselių ir vertikalaus Sobel operatoriaus konvoliucijos būdu apskaičiuotas gradientas.

Toliau Canny algoritmas vykdo pikselių slopinimą esančius toje pačioje gradiento kryptyje, pagal kiekvieno pikselio kaimyną. Pavyzdžiui, jei gradiento kryptis yra horizontali, tuomet pikselis lyginamas su pikselio kairės ir dešinės pusės kaimynais, jei pikselis turi didžiausią reikšmę tuomet lieka nepakeistas, o jei reikšmė mažesnė tuomet pikselis vertė nustatoma 0. Tai reiškia, kad ryškiausios ir aiškiausios briaunos lieka vaizde. Paskutiniame etape taikomas Histerezės slenkstis (angl. *Hysteresis thresholding*), kurio metu pagal pateiktas dvi slenkstines reikšmes nustatoma, ar gautą pikselio gradiento stiprumą galima laikyti briauna. Šiam procesui taikomos dvi slenkstinės reikšmės T_1, T_2. Jei pikselio gradiento stiprumas yra didesnis nei T_2, tuomet jis automatiškai klasifikuojamas kaip briaunos pikselis, jei mažesnis nei T_1, tuomet pikselis yra atmetamas. Tuo tarpu, jei gradiento stiprumas yra tarp $T_1 < T_2$ pikselis laikomas briaunos pikseliu tik tada, jei turi briaunų jungtį su teisingomis briaunomis T_2. Taip užtikrinamas tikslesnis briaunų nustatymas, atmetant nepatikimus ir atsitiktinius objektų kraštus.

Atlikus Canny metodą sekantis žingsnis naudoti Suzuki ir Abe sukurtą objektų kontūrų išskyrimo metodą, kurio pagalba galime gauti rastų kontūrų, objektų detalią informaciją: ilgį, plotį, formą, plotą ir kitas savybes [36]. Informacijos rinkimo procesas vykdomas ieškant kontūro (juodos spalvos pikselio) nuo kairiojo viršutinio vaizdo kampo, judėdamas į dešinę, eilutėmis. Radęs kontūrą atliekamas jo sekimas, pirmasis užtiktas juodas pikselis tampa pradiniu kontūro tašku, tuomet pradedami stebėti jo kaimyniniai 8 pikseliai ir sekami juodi jų kontūrai toliau, kol grįžtama atgal į pradinį tašką. Svarbu paminėti Suzuki ir Abe metodas turi griežtas aprašytas sekimo taisykles, viena iš jų yra paskutinė aptiktų kaimų krypties LNBD (angl. *Last neighbour border direction*) procesas, kuris nurodo paskutinį aptiktą kontūro pikselio poziciją ir kryptį. Aptikęs visus objektus taip surinkdamas prieš tai paminėtas savybes sudaro platų duomenų rinkinį, kuris bus vertingas tolimesniuose mikroplastiko objektų analizės mašininio mokymo etape.

1.5. Objektų aptikimo tikslumo įvertinimas

Visus aptiktus objektus vaizde galime įvertinti apskaičiuojant jų tikslumą panaudojus kontūrų ribojančius keturkampius (angl. *Bounding box*) ir IoU (angl. *Intersection over Union*) metriką. Palyginus originalaus (formulė 1.6, A_t reikšmė), kuris turi faktinį objekto ribojantį stačiakampį, su artimiausiai aptiktu (formulė 1.6, B_p reikšmė) vaizde esančiu keturkampiu, randamas įgyvendintų objektų aptikimo metodų tikslumo įvertinimas IuO, apskaičiuojantis prieš tai paminėtų keturkampių plotų sankirtą, kuri padalijama iš viso abiejų keturkampių ribojamo ploto sąjungos. Žemiau pateiktą IuO vertinimo metrikos 1.6 formulė [8]:

$$IoU = \frac{A_t \cap B_p}{A_t \cup B_p}. \tag{1.6}$$

- $A_t \cap B_p$ - Faktinio objekto ribojančio keturkampio su algoritmo aptiktu ploto sankirta.
- $A_t \cup B_p$ - Faktinio objekto ribojančio keturkampio su algoritmo aptiktu ploto sąjunga.
- IoU - Sankirtos ir sąjungos santykis, kuris parodo ribojančių keturkampių persidengimo efektyvumą.

Ši IoU reikšmė intervalu gaunama nuo 0 (reikšianti visiškai nesutampa) iki 1 (visiškai sutampa), naudojama kaip matavimo metrika objektų aptikimo modeliuose, siekiant įvertinti aptikimo tikslumą (angl. *precision*) ir jautrumą (angl. *recall*). Įvertindami IoU, galime suprasti kaip tiksliai sukurtas modelis lokalizuoja objektus. Nustatę pasirinktą norimą IoU slenksčio vertę, galime sudaryti norimas sąlygas metrikoms išgauti. Reikšmė viršijanti 0.5 gautos faktinių ir aptiktų ribojančių plotų sankirtos ir sąjungos santykio vertei yra priimama kaip teisinga vertė TP (angl. *True Positive*). Jei aptikto ribojančio keturkampio persidengimas mažesnis, tuomet šio stačiakampio negalime vertinti, kaip patikimai susietu su faktiniu objektu, todėl įvertiname kaip klaidingą teigiamą vertę FP (angl. *False Positive*). Jei bent vienam faktiniam objektui nėra aptikto ribojančio keturkampio, tuomet toks atvejis vertinamas klaidinga neigiama verte FN (angl. *False Negative*). Surandant visus šiuos atvejus apibendriname žemiau pateiktuose metrikų 1.7, 1.8, 1.9 formulėse:

$$Precision = \frac{TP}{TP + FP}. \tag{1.7}$$

Tikslumo metrika (1.7) pateikia kiek iš visų identifikuotų objektų yra iš tikrųjų tikrieji aptikti mikroplastiko atvejai.

$$Recall = \frac{TP}{TP + FN}. \tag{1.8}$$

Jautrumo įvertis (1.8) pateikiantis kiek iš visų tikrųjų objekto atvejų teisingai identifikuotos kaip tikrosios mikroplastiko daleles.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}. \tag{1.9}$$

F1 (angl. *F1 score*) įvertis (1.9) teikia balansuotą vertinimą modelio tikslumui ir jautrumui įvertinti.

1.6. Skaitmeninių vaizdų morfologija

Vaizdų apdorojimas taip pat galimas naudojant morfologinį erdvinį pikselių filtravimą. Pasirinktas struktūrinis elementas (stačiakampis, elipsė ir kt. figūros) kartu su lango branduolio dydžiu yra slenkamas per visus vaizdo pikselius, aptikęs objektą pikselis yra pašalinamas arba pridedamas priklausomai nuo naudojamos morfologinės operacijos ir jos veikimo taisyklės. Dažniausiai naudojamos šios žemiau pateiktos keturios funkcijos [13]:

- Erozija (angl. *Erosion*) veikia kai struktūrinis elementas nėra visiškai uždengęs skaitmeninės nuotraukos objekto vaizdo pikselius, tuomet centrinis elemento pikselis nustatomas į 0 (juodą) fono spalvą. Erozijos pagrindinės savybės objekte - šalinti kontūrų ribas, skaidyti į atskiras dalis.

- Plėtimasis (angl. *Dilation*) prideda struktūrinio elemento centre papildomus baltus pikselius prie objekto, kai bent vienas pikselis persidengia su objektu vaizde. Ši savybė naudinga kai norima užpildyti nedidelius tarpus tarp išgautų objektų po binarizavimo metodų, taip sujungiant nutrūkusias objekto dalis.

- Atidarymas (angl. *Opening*) naudojamas, kai norima pašalinti baltus pikselius atskiriant nuo juodo fono. Atidarymas yra erozijos operacija, kuri dažniausiai atliekama prieš plėtimosi operaciją.

- Uždarymas (angl. *Closing*) naudojamas, kai norima užpildyti balta spalva, juodus fono pikselio tarpus baltuose objektuose. Uždarymas yra plėtimosi operacija, kuri dažniausiai atliekama prieš erozijos operaciją tam, kad objekte būtų pašalinti arba sujungti artimi įdubimų pikseliai taip sumažinant objektų deformaciją.

Svarbu atsižvelgti į tai, kad šiame projekte binarizacijos metu objektai sudaro juodą spalvą, o fonas baltą todėl erozija ir plėtimosi funkcijos veiks atvirkštiniu principu. Erozija pridės, o plėtimasis pašalins pikselį.

2. Mašininio mokymosi metodai

Mašininis mokymasis yra dirbtinio intelekto grupės dalis, kuri mokosi ir tobulėja automatizuotai patirties pagrindu. Šis mokymasis yra kitoks kompiuterių programavimo būdas, kadangi pvz. žmogui suprogramuoti pilnai veikiantį kompiuterį visuomet yra sukuriamos apibrėžtos loginės taisyklės, kurios nurodo aiškias veikimo galimybes, savybes. Tačiau mašininiame mokyme programuotojo yra kuriamas tik modelis, kuris gali mokytis iš duomenų rinkinių ir adaptuotis prie naujų duomenų savarankiškai bei gauti rezultatus be žmogaus įsikišimo. Mašininis mokymasis skirstomas į tris pagrindinius tipus [15]:

- Prižiūrimas mokymasis - modelis, kuris mokomas atpažinti duomenų ryšius, kad galėtų teisingai pagal duomenų atributus (angl. *features*) nuspėti nustatytas objektų etiketes (angl. *label*) naujiems anksčiau nematytiems duomenims, kadangi pagal kiekvieną įvesties (angl. *input*) duomenų elementą yra priskiriama konkreti išvesties (angl. *output*) etiketė. Šis būdas naudojamas sprendžiant klasifikavimo (objektų priskyrimas prie klasių) ir regresijos (kintamųjų tarpusavio ryšių vertinimas) užduotis. Šis modelio procesas padeda išmokti, atpažinti duomenų svarbius ryšius leidžiančius teisingai prognozuoti išvesties etiketes naujiems, anksčiau nematytiems duomenims.

- Neprižiūrimas mokymasis - modelis, kuris iš anksto neturi nustatytas duomenų etiketes, o analizė vykdoma remiantis duomenų prielaidomis apie pačių duomenų ypatumus. Pagrindinė šio tipo mokymosi užduotis yra atskleisti paslėptus šablonus, struktūras ar ryšius duomenyse taip suteikiant informaciją, kuri gali būti nepasiekiama prižiūrimo mokymosi metode.

- Skatinamasis mokymasis - modelis, kuris glaudžiai susijęs su grįžtamuoju ryšiu grindžiamu valdymu. Šiame mokyme yra valdomas agentas veikdamas tam tikroje aplinkoje ir stengdamasis optimizuoti savo veiksmus per laiką, kad gautų kuo didesnį atlygį. Tai pasiekiama ne per išankstinį mokymą, o per interakciją su aplinka, kur agentas atlieka veiksmus, stebi aplinkos kitimus ir gauna atlygį arba bausmę už tuos veiksmus. Tuomet agentas adaptuoja savo elgesio strategiją, siekdamas optimizuoti atlygį laikui bėgant. Šis mokymosi būdas yra efektyvus sprendžiant problemas pvz. robotikoje, kur reikia atlikti dinamišką sprendimų priėmimą ar adaptaciją prie kintančios aplinkos.

Šiame mikroplastiko dalelių klasifikavimo užduotyje bus naudojamas būtent prižiūrimasis mokymasis, kurio pritaikymas sprendžia objekto aptikimo, atpažinimo, prognozavimo užduotį naudojant nuotraukų duomenis, o pačio modelio veikimas įvertinamas pagrindinėmis statistinėmis vertėmis: tikslumas (angl. *accuracy*) rodiklis parodantis kaip tiksliai veikia visas modelis, o kiti matavimai, kurie skirti įvertinti atskiros klasifikacijos rinkinį yra tokie patys kaip ir 1.5. skyriuje, (1.7), (1,8), (1.9) formulėse. Žemiau pateiktuose poskyriuose analizuojami projekte naudojami klasifikavimo modeliai.

2.1. Atsitiktinio miško klasifikatorius

Prižiūrimojo mokymo, atsitiktinio miško (angl. *Random forest*) klasifikavimo modelis sudarytas iš sprendimo medžių (angl. *Decision trees*) rinkinio, dažniausiai naudodamas pakavimo (angl. *Bagging* arba pakartojimo agregacijos (angl. *Bootstrap aggregation*) metodo principus [3]. Atsitiktinio miško klasifikatoriaus kiekvieno medžio išsišakojimas apmokomas ne su visais duomenų rinkinio atributais, o tik su atsitiktine duomenų aibe. Tai reiškia, kad kiekvienas duomenų rinkinys

yra išskaidomas į atsitiktines medžio sprendimo dalis, kurio sprendimo medyje yra atsitiktinai pasirenkami atributai pvz. mikroplastiko dalelių apskaičiuoti plotai, kurios yra naudojamos kiekvieno medžio mokymui. Visa tai padeda sumažinti medžių tarpusavio koreliaciją, pagerindami viso modelio tikslumą (angl. *Accuracy*) [4]. Atsitiktinio miško klasifikavimo modelio mokymo procesas yra kompleksiškas, kadangi reikia atrasti tinkamiausius parametrus (angl. *hyperparameter*) su kuriais būtų galima pasiekti efektyviausio modelio veikimą, dažniausiai parametrų parinkimas vykdomas naudojant iš anksto nustatytų parametrų tinkleliu (angl. *hyperparameter grid*). Tinklelis sudaromas iš klasifikatoriaus galimų parametrų kombinacijų, kurios iteracijų principu ieško aukščiausio modelio efektyvumo su specifiniais tyrimo duomenimis. Žemiau pateikti pagrindiniai atsitiktinio miško mokymosi parametrai [28]:

- Medžių skaičius (angl. *Number of trees*) nurodo kiek sprendimų medžių yra sukuriama pagal kurias yra atliekamos modelio prognozės. Didesnis medžių kiekis reiškia geresnį tikslumą, bet pasiekęs tam tikrą ribą tikslumo vertė pradeda nežymiai augti sparčiau didinant skaičiavimo laiką, o ne tikslumą.

- Atsitiktinai parenkamų kandidatinių kintamųjų skaičius (angl. *Number of randomly drawn candidate variables, mtry*) nurodo kiek atributų kintamųjų bus atsitiktinai parenkami kiekviename medžio sprendimų padalijime. Tai padeda sumažinti medžių tarpusavio duomenų koreliacija, kai pasirenkama tik dalinė atributų aibė.

- Mėginių dydis (angl. *Sample size*) nurodo kokia dalis pagrindinio duomenų rinkinio yra naudojama treniruoti kiekvieną atskirą sprendimų medį. Didesnis duomenų kiekis gali sudaryti tarpusavių medžių panašumus, dėl to gali atsirasti didesnė tikimybė modeliui pernelyg gerai prisitaikyti prie mokymo duomenų - persimokymas (angl. *Overfitting*). Persimokymas reiškia, kad modelis pernelyg detalizuodamas mokymo duomenis praranda gebėjimą teisingai prognozuoti naujus, nepažįstamus duomenis. Jei pasirenkamas mažesnis duomenų kiekis treniravimui, medis tampa unikalesniu su išskirtomis specifinėmis savybėmis. Šis unikalumas padeda sukurti medžių įvairovę, tačiau per mažas duomenų rinkinys gali sumažinti kiekvieno atskiro medžio tikslumą, nes mažinant duomenis prarandama informacija apie patį duomenų rinkinį.

- Pakeitimas (angl. *Replacement*) arba dar vadinamas pakartojimu (angl. *bootstrap*) naudojamas kartu su mėginių dydžio parametru kai norima, kad pasirinktas duomenų rinkinys vieno medžio mokymui nebūtų šalinamas iš visos duomenų aibės. Tai reiškia, kad tie patys treniravimo duomenys gali būti dar kartą pasirinkti kitų medžių mokymui. Jei pakartojimas nėra naudojamas tuomet duomenų rinkinys gali būti pasirinktas tik vieną kartą kiekvieno medžio mokymuisi. Tačiau, kaip ir buvo minėta anksčiau, dėl nedidelio mėginių dydžio taip ir su nenaudojamu pakartojimu, per mažas duomenų rinkinys gali sumažinti kiekvieno atskiro medžio prognozavimo naudingumą.

- Mazgo dydis (angl. *Node size*) nurodo minimalų duomenų rinkinio kiekį galutiniame medžio mazge iš kurio nevyksta jokie tolimesni medžio skaidymai ar šakojimai. Nustatant mažesnį mazgo gylį medis išskaidomas į didesnį kiekį šakų, o tai reiškia, kad vykdoma daugiau medžio sprendimų skaidymų iki galutinio mazgo. Šis dydžio parametras gali sukelti persimokymo problemą arba kaip tik pateikti geresnius klasifikavimo modelio rezultatus.

- Skaidymo taisyklė (angl. *Splitting rule*) nurodo kokius atributus skaidyti į skirtingus medžius nusakant skaidymo būdą [40]. Pats skaidymas grindžiamas šiomis pagrindinėmis tai-

syklėmis: Gini nevienodumo indeksu (angl. *Gini impurity index*) ir Entropija (angl. *Entropy*). Gini indeksas, remdamiesi formule (2.1), siekia minimalizuoti tikimybę $\hat{p}_{mk}$ (kiekvienos klasės K atributų proporciją konkrečiame mazge), kad atsitiktinai pasirinktas atributas būtų klasifikuojamas neteisingai. Tai reiškia, kad algoritmas ieško kiekvieno atributo skaidymo varianto, kuris duotų mažiausią Gini nevienodumo reikšmę.

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}). \tag{2.1}$$

Entropija (formulė 2.2) kaip ir Gini skaidymo taisyklė kiekviename sprendimų medžio mazge ieško mažiausios E (entropija) nevienodumo vertės, tačiau jie apskaičiuojami skirtingais būdais, o tai gali duoti šiek tiek skirtingus rezultatus atsižvelgiant į konkrečius duomenis.

$$E = -\sum_{k=1}^{K} \hat{p}_{mk} \log(\hat{p}_{mk}). \tag{2.2}$$

Taip pat dar yra taikomi papildomi *scikit-learn* bibliotekos atsitiktinio miško parametrų sprendimai: didžiausias medžių gylis (angl. *max_depth*), mažiausias skaidymui reikalingų mėginių skaičius (angl. *min_samples_split*) ir mažiausias galutinio mazgo turimų mėginių skaičius (angl. *min_samples_leaf*). Didžiausias gylio parametras nurodo medžių skaidymo limitą, tai padeda kontroliuoti per tikslų modelio prisitaikymą išvengiant persimokymo problemos. Mažiausią skaidymo mėginių skaičius nurodo ties kokiu mažiausiu turimų savybių (mėginių) kiekiu bus atliekamas skaidymas, o galutinis mazgo mėginių limitas užtikrina, kad po skaidymo sudarytuose paskutiniuose medžių taškuose būtų pakankamai likę mėginių skaičius.

Atsitiktinio miško klasifikatoriuje pasitelkę parametrų tinklelio sprendimą galime rasti optimaliausius parametrus, kurie užtikrina efektyvų balansą tarp persimokytų ir nepakankamai apmokytų (angl. *underfitting*) medžių sprendinių. Šis balansas leidžia sukurti patikimą modelį, kuris gebėtų efektyviai atpažinti klases su įvairiais duomenų rinkiniais.

2.1.1. Savybių svarbos nustatymo metodas

Taikant išgautų atributų (5.1. skyrius) svarbos metodą pagal vidutinį sumažėjimo nevienodumo parametrą (angl. *Mean decrease in impurity, MDI*) atsitiktinio miško klasifikatoriuje pateikiamos optimaliausios aptiktos objekto savybės, kurios daugiausiai arba mažiausiai naudingos modelio sprendimų priėmimo procesuose, kadangi nereikšmingos ar mažai informatyvios savybės trukdo modeliui atrasti svarbius duomenų ryšius, dėl kurių modelis negali pasiekti efektyviausių rezultatų. Šis metodas taikomas kartu su skaidymo taisyklėmis (entropija, gini metodai) tačiau pats MDI skaičiavimas nėra priklausomas nuo to, jis tik naudojasi nevienodumo vertėmis randant nevienodumo sumažėjimo vertę $\Delta I(t)$. Žemiau pateiktos formulės (2.3) parodančios kaip apskaičiuojamos visos atsitiktinio miško gaunamos skirtingos atributų MDI vertės:

$$MDI(k,T) = \sum_{t \in I(T), v(t)=k} \frac{N_n(t)}{n} \Delta I(t) \quad \text{ir} \quad MDI(k) = \frac{1}{n_{\text{tree}}} \sum_{s=1}^{n_{\text{tree}}} MDI(k, T_s). \tag{2.3}$$

- $MDI(k,T)$ - vieno medžio T, konkrečios k (atributo indeksas) savybės MDI vertė.
- t - medžio mazgo skaičius.

- $I(T)$ - viso miško vidinių mazgų aibė medyje T.
- $v(t) = k$ - nurodoma, kad mazgas t padalijamas naudojant savybę k.
- $N_n(t)$ - duomenų kiekis patenkantis į mazgą t.
- $\Delta I(t)$ - nevienodumo sumažėjimas mazge t, apskaičiuojamas tėvinio mazgo nevienodumo vertės ir sumos abiejų vaikų mazgų nevienodumo skirtumui.
- n - duomenų kiekio skaičius sprendimų medyje.
- n_{tree} - medžių skaičius miške.
- $MDI(k)$ - vidutinė viso atsitiktinio miško nevienodumo mažėjimo reikšmė pagal kiekvieną unikalią k savybę.

Pagrindinis skaičiavimas atliekamas sumuojant nevienodumo mažėjimo vertes $MDI(k)$ visose specifiniuose medžio vidiniuose mazguose, kurie yra padalijami pagal skirtingas savybes k. Kiekvienas toks mazgas prisideda prie bendros savybės k svarbos medyje, priklausomai nuo to, koks duomenų kiekis $N_n(t)$ patenka į mazgą ir kokia yra to nevienodumo sumažėjimo vertė $\Delta I(t)$ tame mazge po skaidymo. Visam šiam skaičiavimui naudojami $t \in I(t)$ visi medžio T vidiniai mazgai užtikrinant sąlyga $v(t) = k$, kad į skaičiavimą būtų įtraukti tik tie mazgai, kurie padalinti naudojant savybę k [16].

2.2. KNN klasifikatorius

Šio k - artimiausių kaimynų (angl. *K - nearest neighbours*) modelio veikimas yra paprastesnis nei atsitiktinio miško klasifikatoriaus. Pats modelis išsivystė iš artimiausių kaimynų (angl. *Nearest neighbours*) algoritmo kuris naudoja k reikšmę sprendimams priimti [9]. Šis parametras ir yra vadinamas kaimynų skaičiumi k, kuris nurodo kiek artimiausių taškų iš mokymo duomenų rinkinio bus naudojama klasifikavimo sprendimui priimti. Prieš priimant sprendimą algoritmas pirmiausia apskaičiuoja kiekvieno mokymo atributo esančio taško atstumą, kuris randamas pagal Euklido arba Manheteno formules:

- Euklido atstumas $d(p, q)$ apskaičiuojamas pagal žemiau pateiktą formulę (2.4). Atributų taškų koordinatės p ir q n - matėje erdvėje apskaičiuojamos pagal rastų taškų skirtumus tarp taškų koordinačių kvadratų sumos ir traukiamos kvadratinės šaknies.

$$d(p, q) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}. \tag{2.4}$$

- Manheteno atstumas $d(p, q)$ apskaičiuojamas pagal žemiau pateiktą formulę (2.5). Šis atstumas apskaičiuojamas tarp p ir q atributų koordinačių taškų n - matėje erdvėje sumuojant koordinačių skirtumų absoliutines vertes. Skirtingai nei Euklido atstumas, kuris matuoja tiesioginį atstumą, Manheteno principas yra matuoti tik horizontaliomis ir vertikaliomis linijomis.

$$d(p, q) = \sum_{i=1}^{n} |q_i - p_i|. \tag{2.5}$$

Kitas svarbus parametras - svoris, kuris suteikia galimybę kiekvienam k artimiausiam duomenų kaimynui įgyti svorį vertinant kokiai klasei priklauso naujas mokymo duomenų taškas. Tai reiškia, kad kiekvienas kaimynas priklausomai nuo atstumo įgaus didesnę arba mažesnę vertę sprendimui priimti, kokia tai klasė. Parametrų tinklelio sprendimas naudingas randant optimaliausius parametrus šiam KNN klasifikatoriui [35].

2.2.1. Atributų svarbos nustatymo ReliefF metodas

Pats KNN klasifikatorius neturi būdo įvertinti atributų svarbumo, kaip atsitiktinis miškas skaičiuojant vidutinį sumažėjimo nevienodumo dydį. Tačiau pasitelkus išorinį metodą ReliefF galima įvertinti atributų svarbumą pasirinktame klasifikatoriuje. Šiam algoritmui vykdant pasirenkamas atsitiktinis teigiamos klasės atributas iš duomenų rinkinio, tuomet ieškomi to duomens artimiausi kaimynai tos pačios klasės, kurie dar vadinami pataikymai (angl. *hits*), kitų klasių rasti artimiausi kaimynai vadinami nepataikymais (angl. *misses*). Gavus galutinius rezultatus, ReliefF metodas vertina kiekvieno atributo skirtumą tarp pataikytų ir nepataikytų duomenų. Jei tam tikras atributas labai skiriasi tarp šių dviejų grupių, tuomet atributas yra vertinamas kaip svarbus. ReliefF analizuoja, kaip kiekvienas atributas skiriasi tarp pataikymų (kaimynų teigiamoje klasėje) ir nepataikymų (kaimynų neigiamoje klasėje). Šis vertinimas atliekamas pagal atstumo skirtumus, pavyzdžiui, jeigu atributas teigiamoje klasėje ir pataikyme yra labai panašūs, bet tarp teigiamos klasės ir nepataikymo skiriasi, tai parodo, kad teigiamos klasės atributas yra svarbus klasifikacijos užduočiai. Artimiausi kaimynai randami pasitelkus KNN klasifikatoriaus atstumo matavimo metodo principus [37] [38].

2.3. AdaBoost klasifikatorius

Adaptyvaus stiprinimo mokymo algoritmas (angl. *AdaBoost*) vykdomas jungiant priklausomus vienas nuo kito sprendimo kelmus (angl. *Decision stumps*). Sprendimo kelmas - supaprastintas sprendimų medžių variantas, turintis tik vieną mazgą. Kiekvienas atskiras kelmas dar vadinamas silpnuoju klasifikatoriumi. Pirmojo klasifikatoriaus rezultatai nėra tikslūs, kadangi daro pernelyg daug klaidų, todėl jo efektyvumas įvertinamas mažu svoriu, o netinkamai klasifikuotiems duomenims (mėginiams) priskiriami didesni svoriai, tai reiškia, kad sekančiame klasifikatoriuje algoritmas skirs daugiau dėmesio būtent klaidingai atpažintoms klasėms. Sekantis medis mokomas gautais duomenų svoriais ir jei jo gautas modelio tikslumas optimalesnis jam priskiriamas didesnis svoris, todėl modelis mokomas kol galiausiai pasiekiamas galutinis iteracijų kiekis arba kai modelis daugiau nesugeba pagerinti modelio rezultatų. Galutinis sprendimas vadinamas stipriuoju klasifikatoriumi, kuris sudaromas iš kiekvieno silpnojo klasifikatoriaus priskirtų išvesties rezultatų (kiekvieno sprendimų kelmo priskirtos prognozuojamos klasės - teigiamos 1 arba neigiamos -1 reikšmės, kurias modelis nustato pagal teisingas faktinės klasės savybes) sandaugos iš to pačio sprendimo kelmo turimo svorio, tuomet šie svoriniai rezultatai susumuojami tarpusavyje ir modelis galutinai nusprendžia kuriai klasei priklauso prognozuojamas objektas [14].

2.4. XGBoost klasifikatorius

Ekstremalus gradiento stiprinimo (angl. *Extreme gradient boosting*) klasifikatorius taip pat kaip ir adaptyvus stiprinimo metodas veikia jungdamas sprendimų medžius į vieną stiprųjį klasifikatorių. Tačiau XGBoost metodas naudoja gradientinio nusileidimo sprendimą. Kiekvienas

naujai pridedamas medis (vidiniame modelyje) yra mokomas, kaip geriausiai pakeisti modelį, kad sumažėtų nuostolių funkcijos (angl. *Loss function*) vertė, taip pagerinant modelio mokymų rezultatus. Šiame tyrime naudojama logistinė nuostolių funkcija, kurios pagrindas tikimybių vertinimas. Ekstremalaus gradiento stiprinimo modelyje šis būdas įvertina modelio prognozavimo tikslumą. Modelis ne tik atpažįsta, ar duomenys priklauso tam tikrai klasei, bet ir įvertina, kaip tikėtina, kad šis prognozavimas yra teisingas. Ši funkcija skaičiuodama nuostolį už kiekvieną modelio atliktą prognozę įvertina kaip toli yra nuo faktinių duomenų klasės. Jei modelio įvestis priklauso teigiamai klasei, bet iš tikrųjų ji priklauso neigiamai tuomet logistinės nuostolių funkcijos vertė bus didelė [20]. Priklausomai nuo pasirinktos nuostolių funkcijos gradiento stiprinimo modelyje siekiama nustatyti, kaip skiriasi modelio prognozės nuo faktinių duomenų, kur kiekvienoje mokymo iteracijoje apskaičiuojamas jų skirtumas. Tuomet naudojamas gradientas (vektorius), kuris nurodo, kaip nuostolių funkcijos vertė keičiasi, koreguojant modelio parametrus. Taip siekiama padidinti bendrą modelio našumą, kiekvieną kartą mažinant nuostolių funkcijos vertę. Ekstremalus gradiento stiprinimo modelis yra įtraukęs reguliarizacijos būdus į mokymo algoritmo nuostolių funkciją skatinant modelį rinktis paprastesnes struktūras ir mažesnius svorius, taip sumažinant modelio persimokymo riziką ir vengiant gradiento lokalaus minimumo (įdubimų). Modelio reguliarizacijos parametrai: L2 λ (angl. *reg_lambda*) ir L1 α (angl. *reg_alpha*). Reguliarizacija L1 dar vadinama Lasso reguliarizacija, kuri klasifikatoriuje naudojama atributų atrankai. Sumažindamas tam atributų svorių vertę iki nulio, efektyviai pašalina mažiau svarbias savybių duomenis iš modelio (mažiau svarbios savybės yra mažą svorį turintys atributai). Parametras α algoritme yra nustatytas kaip numatytoji reikšmė 0, tokia vertė modelio neformuos į retesnę struktūrą. Reguliarizacija L2 dar vadinama Ridge reguliarizacija, kuri mažina visas atributų svorių vertes, bet nepasiekia absoliučios nulinės reikšmės. Tai leidžia išlaikyti modelio stabilumą sumažindamas persimokymo riziką. Kitas *scikit-learn* bibliotekos parametras, tai γ minimalus medžių mazgų skaidymo skaičius (angl. *min_split_loss*) reguliuoja medžių mazgų sudėtingumą, didinant reikšmę, padidinama nuostolio vertė už kiekvieną papildomą mazgą medyje. Tai reiškia, kad modelis yra skatinamas naudoti paprastesnes medžių struktūras su mažiau mazgų. Kitas svarbus parametras - λ, kuris reguliuoja mazgų svorių dydį (įvertis galutiniam modelio sprendimui), dėl per didelių svorių yra padidinama nuostolių funkcijos vertė, taip skatinant modelį naudoti mažesnius svorius, kurie leidžia sumažinti modelio kompleksiškumą [7].

Praktinėje dalyje taikyti papildomi XGBoost galimi parametrai, tai mokymo duomenų dalis (angl. *subsample*), savybių mokymo dalis (angl. *colsample_bytree*) ir svorių normalizacija (angl. *normalize_type*). Parametras *subsample* modelyje nustato, kokia procentinė dalis visų mokymo duomenų bus naudojami kiekvieno medžio mokymui. Kitas parametras *colsample_bytree* veikia panašiai, tačiau nustato, kokia procentinė dalis savybių bus naudojami kiekvieno medžio mokymui. Normalizacijos parametrą sudaro dvi pagrindinės reikšmės tai medžio (angl. *tree*) ir miško (angl. *forest*). Kai pasirenkamas medžio parametras, tuomet kiekvieno medžio svoris yra normalizuojamas individualiai, leidžiant kiekvienam medžiui turėti atskirą mokymo svorį, o miško parametras nurodo, kad visų medžių svoriai normalizuojami kartu kaip viena visuma, užtikrinant subalansuotą visų medžių įtaką modelio sprendimui. Šis parametras svarbus, kadangi padeda kontroliuoti, kaip skirtingi modelio komponentai prisideda prie galutinio rezultato.

2.5. Sunkių neigiamų pavyzdžių metodas

Sunkūs neigiami pavyzdžiai (angl. *Hard negative mining*) yra modelio neteisingai įvertinti duomenys, kai neigiamai prognozuotas objektas (neteisingai klasei priklausantis) yra nustatomas

kaip teigiamas (priklausantis teisingai klasei). Įgyvendinant tokį metodą pirmiausia vykdomas pradinis klasifikatoriaus modelio mokymas su dvejomis nesubalansuotomis klasėmis: ne mikroplastikas ir mikroplastikas. Tuomet validavimo rinkinyje, reguliariai įvertinant modelio efektyvumą, identifikuojami didesnės klasės sunkūs neigiami pavyzdžiai (objektai labai panašūs į teigiamą klasę), kuriuos modelis galimai neteisingai klasifikavo ir jie yra įtraukiami į naują duomenų mokymo rinkinį, kur modelis mokomas iš naujo. Šis metodas vykdomas iteracinių mokymų pagrindu, kadangi kiekvienoje iteracijoje modelis susiduria su naujais neteisingai prognozuotomis klasių atvejais todėl modelis yra mokomas iš naujo su naujai atrinktais sunkiaisiais neigiamais pavyzdžiais siekiant gauti efektyviausią rezultatą. Per dažnas modelio mokymas gali sukelti duomenų persimokymą, tačiau tai nereiškia, kad paskutiniai, iki persimokymo esamos iteracijos rezultatai yra optimalūs, kadangi geriausi rezultatai (tikslumas, jautrumas, F1) gali būti gaunami jau pirmosios iteracijos metu, todėl svarbu stebėti visus mokymo etapus [41] [22].

2.6. Sintetinis klasės mažumos didinimo metodas

Sintetinės mažumos duomenų didinimo metodas (angl. *Synthetic minority over - sampling technique, SMOTE*), kuris dirbtinai didina klasės duomenų rinkinį. Šis metodas remiasi KNN algoritmu, kur pačioje pradžioje atsitiktinai pasirenkamas mažesnės klasės vienas atributo mėginys tam, kad identifikuotų k artimiausius tos pačios mažumos klasės taškus. Toliau iš šių k artimiausių kaimynų atsitiktinai pasirenkamas sekantis kaimynas, ir tarp šio pasirinkto mėginio bei atsitiktinai pasirinkto kaimyno sukuriami nauji duomenys. Šis procesas kartojamas tol kol pasieks pasirinktą klasės balansą, generuojant šiuos dirbtinius duomenis, kurie padeda pasiekti geresnį klasės pasiskirstymo balansą duomenų rinkinyje [6].

3. Susijusių darbų apžvalga

3.1. Automatinis mikroplastiko dalelių skaičiavimas ir klasifikavimas

Literatūroje mikroplastiko dalelių išskyrimas nuo fono yra viena iš pagrindinių segmentavimo dalių. Tyrinėtame darbe vaizdai ganami naudojant skaitytuvą tai reiškia, kad nuotraukos fono spalva išlieka ganėtinai šviesi, o dalelės lieka tamsesnės. Siekiant visiškai išskirti objektus naudojami Otsu ir Sauvola binarizavimo metodai. Sauvola metodas efektyvesnis šiame tyrime, kadangi jis tiksliau segmentavo visus esančius mikroplastiko fragmentus. Otsu nesugebėjo tinkamai aptikti linijinių objektų, kurios objektus skaidydavo į atskiras dalis. Toliau siekiant klasifikuoti mikroplastiką buvo panaudoti penki skirtingi mašininio mokymosi algoritmai: KNN , C4.5 (angl. *Decision Tree*), atsitiktinis miškas (angl. *RF*), SVM (angl. *Support Vector Machine*), AdaBoost (angl. *Adaptive boosting).* Pagal eksperimentinius rezultatus, RF su ReliefF metodo integracija leido pasiekti aukštą klasifikavimo tikslumą - 96,6 %. ReliefF metodas naudojamas atrasti reikšmingiausias savybes, kurios efektyviausiai padeda atskirti mikroplastiko daleles. Straipsnyje nurodyta, kad pasitelkus binarizacijos metodus iš kiekvienos aptiktos dalelės buvo išgaunamos tam tikros spalvinės ir geometrinės savybės [17].

3.2. Mikroplastiko skaičiavimo ir klasifikavimo automatizuota programinė įranga

Šiame moksliniame straipsnyje tie patys, kaip 3.1 poskyrio autoriai, pratęsę tyrimą sukūrė programinę įrangą SMACC (angl. *System for microplastics automatic counting and classification*). Tai atvirojo kodo programinė įranga, kuri sukurta siekiant automatizuoti mikroplastikos dalelių klasifikavimo ir kiekybinės analizės procesą. Tyrime segmentavimas buvo atliekamas naudojant Sauvola metodą, o mikroplastiko klasifikavimas atliekamas kaskadiniu metodu, kadangi realiu laiku veikiančiose metoduose, greitis yra labiau vertinamas nei tikslumas. Kaskadinį klasifikavimą sudaro: KNN, C4.5, atsitiktinis miško ir SVM klasifikatoriai. Kaskadinė klasifikavimo strategija pasirinkta, dėl skirtingų mikroplastiko dalelių tipų: linijų, granulių ar kitų objekto fragmentų. Iš pradžių suklasifikuojamos paprastesnės dalelės ir taip klasifikavimo hierarchijos principu keičiamas klasifikavimo metodas ieškant kito sudėtingesnės mokroplastiko figūros rūšies. Kaskadinis modelio naudingumas pasiekė 91,1 % tikslumo rezultatą. Šis metodas leido efektyviau naudoti klasifikavimo resursus, ypač kai klasės gali būti lengvai atskirtos remiantis mažesnių atributų rinkiniu.

Darbe taip pat atliekamas klasifikavimas naudojant konvoliucinį neuroninį tinklą, tačiau dėl duomenų trūkumo buvo atlikta vaizdų augmentacija, taip sukuriant daugiau duomenų pavyzdžių. Tuomet naudojamas duomenų rinkinys padalinamas į 10 atsitiktinių rinkinių pasitelkiant 10 kryžminį validavimą (angl. *Cross validation*), šis metodo modelis mokamas naudojant 9 iš tų esamų duomenų rinkinių ir testuojamas su likusiu vienu. Šitaip kartojamas 10 kartų, kiekvieną kartą su skirtingu testavimo skirtu daliniu duomenų rinkiniu, taip įvertinant neuroninių tinklų veikimą. Rezultatai parodė aukštą tikslumą - 97,4 %, kuris viršijo prieš tai paminėtų klasifikatorių rezultatus. Tačiau eksperimento sąlygos nebuvo visiškai lygios todėl tai liko tik perspektyva šiai tyrimo krypčiai ir į SMACC programą nebuvo įtrauktas šis sprendimas [19].

3.3. Giliojo mokymo metodas automatiniam mikroplastiko skaičiavimui ir klasifikavimui

Mokslininkų mikroplastiko dalelių nuotraukos sudaromos baltame fone naudojant kelis skaitmeninius fotoparatus. Pirmojoje pagrindinėjė projekto dalyje skirtas dėmesys dalelių segmentavimui naudojant U-Net konvoliucinį neuroninį tinklą, kurios ypatumas leidžia gauti aukštos kokybės segmentavimo rezultatus. Antroje projekto dalyje naudojant VGG16 (angl. *Visual geometry group*) neuroninį tinklą mikroplastikas buvo klasifikuojamas į tris tipus: fragmentus, granules ir linijas. VGG16 konvoliucinių neuroninių tinklų modelis turi 16 sluoksnių, tai leidžia išgauti abstrakčias ir tikslias nuotraukos savybes, kas yra būtent svarbu klasifikavimui. Svarbu paminėti, kad šie konvoliucinių sluoksnių svoriai buvo inicijuoti iš anksto jau apmokytų (angl. *ImageNet*) vaizdų duomenų bazės, tai padėjo darbo autoriams sumažinti reikiamą mokymo duomenų kiekį. Tyrimo rezultatai parodė, kad U-Net segmentavimas pranašesnis už Sauvola segmentavimą, o galutinis klasifikavimo tikslumas siekė 98,11 % [18].

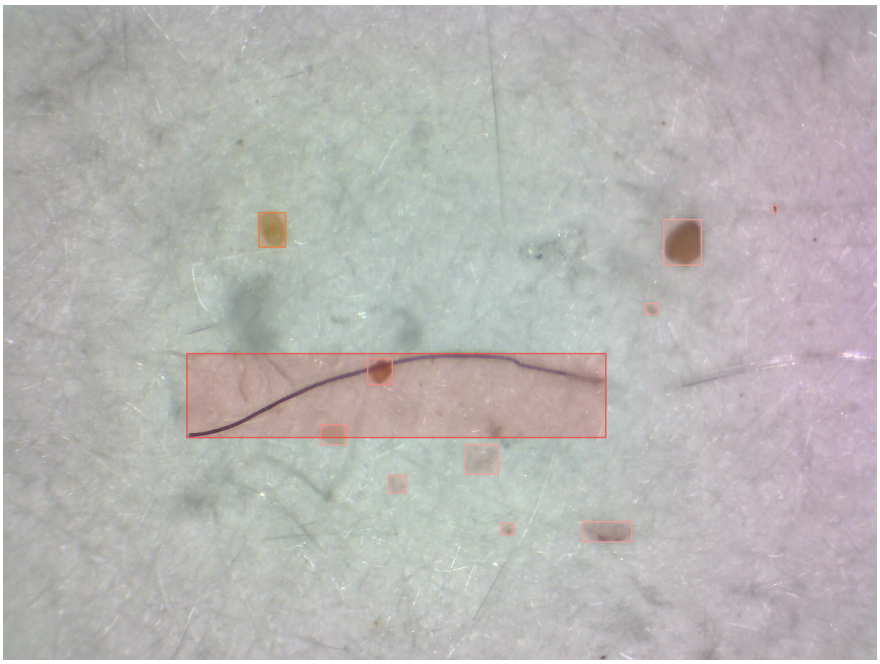
3.4. Automatinis mikroplastiko kiekio nustatymas ir klasifikavimas taikant gilųjį mokymąsi

Ištirtame projekte vykdomas mikroplastiko dalelių automatinis kiekybės skaičiavimas ir klasifikavimas pasitelkus skenuojančios elektroninės mikroskopijos (angl. *SEM*) vaizdus. Projekte naudotas duomenų rinkinys susideda iš 237 SEM nuotraukų, kuriose vaizduojamos mikroplastiko dalelės. Dalelių segmentavimas vykdomas taikant giliojo mokymo modelius U-Net ir *MultiResUNet*. Ši pastaroji U-Net modelio modifikacija, pritaikyta geriau tvarkyti vaizdų su skirtingo lygio detalėmis ir tekstūromis. *MultiResUNet* įtraukia tam tikrus blokus, kurie leidžia neuroniniui tinklui efektyviau dirbti su įvairaus dydžio vaizdo savybėmis bei tikslingiau atpažinti ir segmentuoti objektus su skirtingais dalelių dydžiais ir jų formomis. Projekte dalelių klasifikavimui taikomas VGG16 konvoliucininis neuroninis tinklas mikroplastiko formų klasifikavimui, kuris pasiekia 98.33 % tikslumą [33].

4. Mikroplastiko dalelių segmentavimo tyrimas

4.1. Duomenų aprašymas

Projekto metu vykdomi eksperimentai naudojantis 405 anotuotomis mikroplastiko *jpg* formato nuotraukomis, kurių pikselių raiška 24 bitų gylyje yra (spalvų diapazonas raudona, žalia, mėlyna) 2592 x 1944 vaizdo dydžio. Patį rinkinį išviso sudaro sudaro 852 išskirti mikroplastiko objektai, kurios suskirstytos į tris grupes: 702 fragmentai, 129 gijos ir 21 granulė. Šis duomenų rinkinys gautas iš fizinių ir technologijos mokslų centro (FTMC) aplinkotyros skyriaus mokslo darbuotojos dr. Ievos Uogintės. Žemiau pateikta nuotrauka (1 pav.) (iš šiam projekto tyrimams gauto vaizdų rinkinio) su 3 atvaizduojamomis mikroplastiko klasėmis: fragmentas, gija ir granulė. Mikroskopijos vaizdai nėra vienalyčiai, juose matomas netik ryškiai išsiskiriantis mikroplastikas, bet ir fone neryškios pilkos dėmės, kurias negalėtume vertinti kaip mikroplastiko objektais, kadangi tai dažniausiai yra mikroskopu matomos nuosėdos, kurie pateikia vaizdą su tam tikromis neryškiomis dėmėmis. Visi mikroskopijos nuotraukuose esantys mikroplastikai yra sužymėti ribojančiais spalvotais stačiakampiais, išskirti skirtingomis spalvomis. Kiekviena spalva išskiria tris mikroplastiko klases: gijas, fragmentus ir granules. Visos nuotraukų rinkinyje pavaizduotos mikroplastiko dalelės anotuotos dr. Ievos Uogintės naudojant MakeSense.ai įrankiu, kurias [34], kuriame suklasifikuotas mikroplastikas išsaugotos JSON duomenų formatu (angl. *JavaScript Object Notation*) [27].



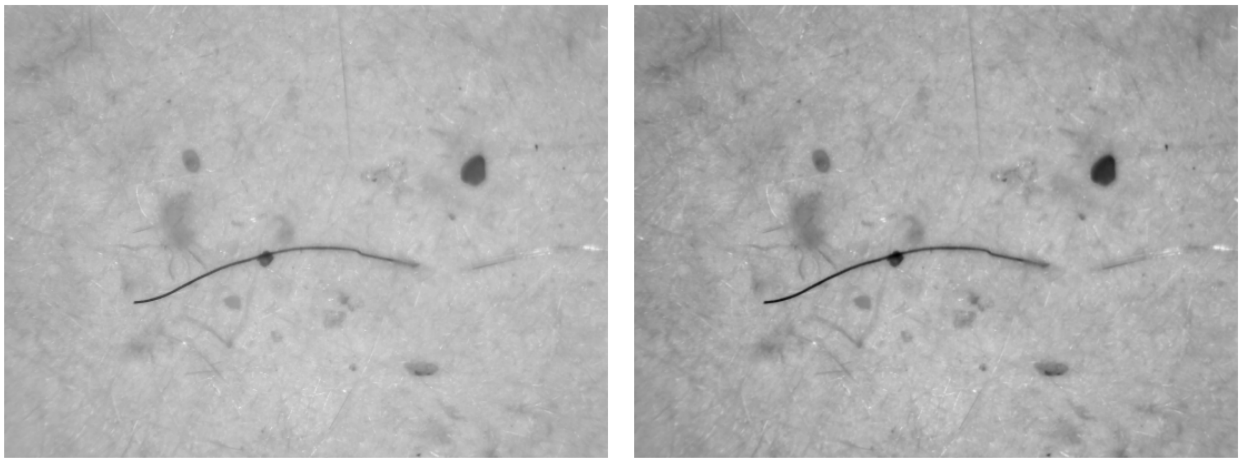
1 pav. Mikroskopo vaizdas su išskirtomis faktinėmis mikroplastikų objektais.

Tyrimams naudojama *Google Colab* [2] platforma, kuri suteikia galimybę Python programavimo kalba [39] atlikti eksperimentinius tyrimus, kompleksiniams segmentavimo ir mašininio mokymo uždaviniams spręsti. Ši platforma turi galingus skaičiavimo išteklius padedančius apdoroti didelius duomenų rinkinius. Be to, nesudėtingai integruojasi su *Google drive* duomenų saugykla, užtikrindama efektyvų duomenų valdymą ir saugojimą. Dėl šios integracijos, projektui gautos mikroplastiko dalelių nuotraukos yra patogiai, greitai pasiekiamos, eliminuojant poreikį nuolat (iš naujo paleidus *Google Colab*) įkelti vaizdus į platformą.

4.2. Globalaus ir adaptyvaus binarizavimo metodų palyginimas

Pirmoje eksperimento dalyje, įvertinant rezultatus, palyginami gana plačiai žinomi ir naudojami vaizdų segmentavimui skirti binarizavimo metodai, Otsu ir Sauvola. Nors 3.1. skyriuje literatūroje pateikiama, kad mikroplastiko dalelių atpažinimo užduotyje Sauvola metodas yra efektyvesnis, tačiau šiame tyrime naudojant kitokį mikroplastiko vaizdų duomenų rinkinį (sudarytą iš mikroskopo nuotraukų) galime gauti kitokius rezultatus būtent konkretaus eksperimento sąlygomis.

Prieš binarizavimą atliekamas vaizdo apdorojimas, kurio eiliškumas grindžiamas vaizdo paruošimo teorinėje dalyje (1.1. skyrius). Projektui vykdyti naudojama *OpenCV* atvirojo kodo biblioteka [23], kuri yra pritaikyta dirbti su vaizdo apdorojimu ir jo analize. Visą pradinį vaizdų rinkinį, be mikroplastiko spalvotų ribojančių keturkampių kaip pavaizduota (1 pav.), konvertuojame į pilkos skalės nuotraukas, bei po to atliekame Gauso triukšmo glodinimo metodą. Triukšmo filtravime naudojamas 5x5 branduolys, kurio nuokrypio konstanta yra apskaičiuojama automatiškai priklausomai nuo pasirinkto branduolio dydžio pagal *OpenCV* biblioteką. Iš žemiau pateiktų nuotraukų (2 pav.) galime įvertinti, kad mikroplastiko dalelės po atlikto Gauso vaizdo glodinimo tapo ryškesnės (2 pav. (b)), o tai leis gauti aiškesnius ir tikslesnius binarizavimo rezultatus.



(a) Pilkos skalės nuotrauka su triukšmu. (b) Vaizdas po triukšmo glodinimo.

2 pav. Mikroskopo nuotrauka pilkoje skalėje prieš ir po triukšmo filtravimo.

4.2.1. Otsu binarizavimo taikymas

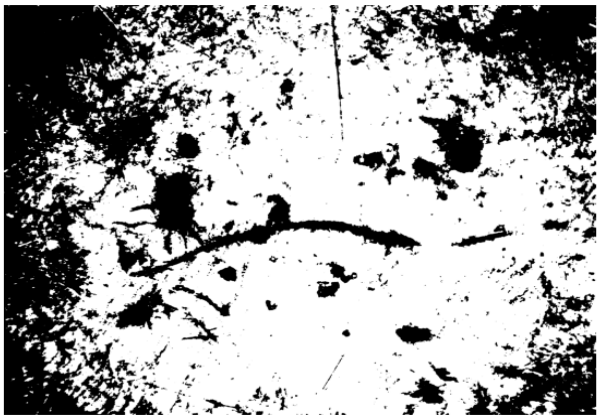
Galime pamatyti, kad nuotraukos (1 pav.) kraštuose yra patamsėjimai, centre vaizdas šviesesnis, o toks veiksnys yra pasikartojantis visame vaizdų rinkinyje. Šis faktorius turi labai reikšmingos įtakos binarizavimo procesui, kadangi tai sumažina esamą kontrastą tarp objektų ir fono. Tokiu atveju, naudojant Otsu metodą, mikroplastikų dalelės, ypač esančios vaizdo kraštuose, yra sunkiai atskiriamos nuo fono dėl nepakankamo šviesos skirtumo, kurį galime pamatyti 3 pav. (a) kairės pusės nuotraukoje.

Be papildomų vaizdo šviesumo ir kontrasto korekcijų Otsu binarizavimas išlieka netikslus. Todėl nuspręsta pasinaudoti *OpenCV* bibliotekoje esama (4.1.) tiesinės transformacijos lygtimi:

$$G(x, y) = \alpha f(x, y) + \beta. \tag{4.1}$$

Ši lygtis parodo, kaip nuo skirtingų α ir β koeficientų pradinis vaizdo pikselis $f(x, y)$ yra keičiamas, kad būtų gautas atitinkamo šviesumo ir kontrasto galutinė vaizdo išvesties pikselio ($G(x, y)$

koordinatėse) reikšmė. Koeficientas α kontroliuoja vaizdo kontrastą, reguliuodamas pikselių atspalvių skirtumą tarp šviesių ir tamsių objektų, o β koeficientas modifikuoja kiekvieno vaizdo pikselio šviesumo intensyvumą. Pritaikius šį metodą galime pamatyti (3 pav. (b) nuotraukoje), atsitiktinai parinktomis α ir β vertėmis, kad nuotraukos kraštuose nebelieka juodo atspalvio, vaizdo fonas tampa visiškai baltas, išskirdamas tik keletą juodų dalių.

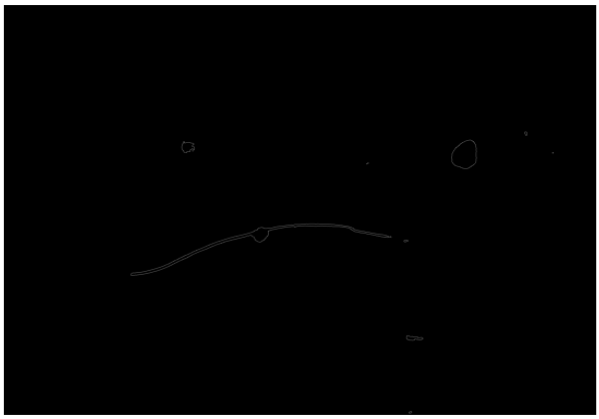


(a) Gauta slenkstinė vertė $T = 179$, nenaudojant koeficientų.

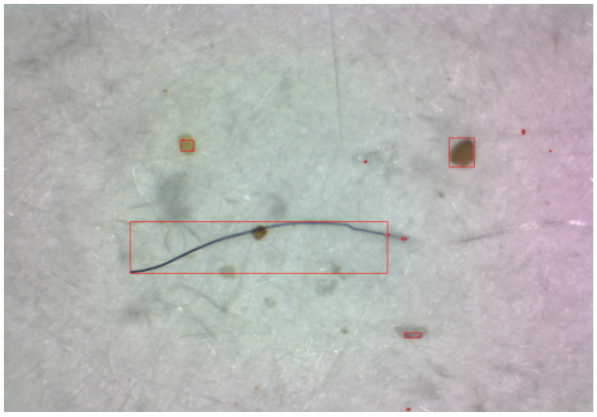
(b) Gauta slenkstinė vertė $T = 223$, su $\alpha = 1$ ir $\beta = 90$ koeficientais.

3 pav. Otsu binarizavimo metodo gautos skirtingos slenkstinės vertės, priklausomai nuo vaizdo šviesumo ir kontrasto koeficientų.

Sekančiame tyrimo etape atliekamas objektų kontūrų išskyrimas (1.4. skyrius), naudojant *OpenCV* bibliotekoje esančius Canny ir Suzuki metodų sprendimus. Tai leidžia nustatyti galimas mikroplastiko daleles ir apibrėžti jas apgaubiančiais keturkampiais. Visus binarizacijos metu išryškintas daleles, Canny metodas išskiria objektų briaunas ir pateikia juodo fono nuotrauką su baltos spalvos kontūrais (4 pav. (a)), o Suzuki metodas radęs to objekto geometrinius parametrus suteikia galimybę nubraižyti galimam mikroplastikui ribojantį raudonų linijų keturkampį (4 pav. (b)).



(a) Išskirti objektų kontūrai juodame fone panaudojant Canny metodą.

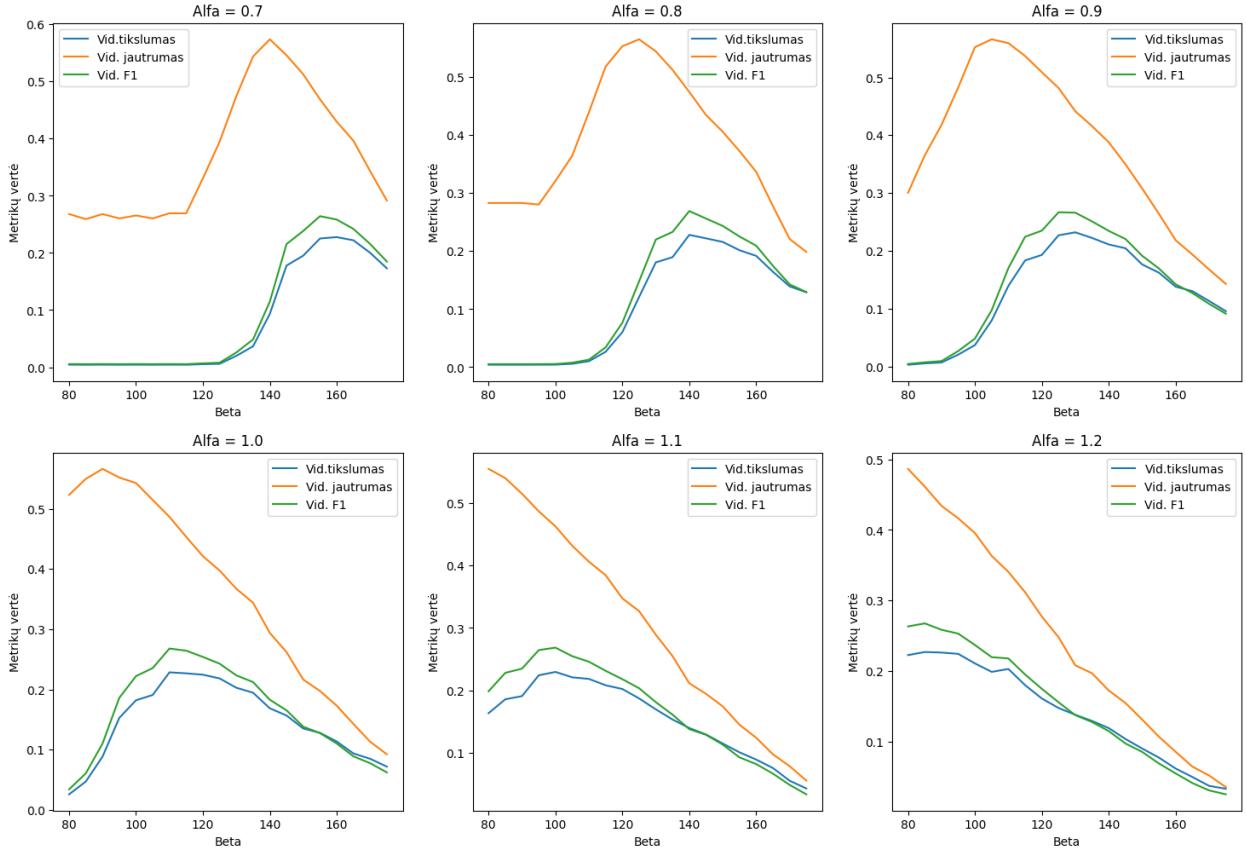


(b) Pažymėti objektų kontūrai panaudojant Suzuki metodą.

4 pav. Įgyvendintas Canny ir Suzuki metodas, kurių pagalba išskiriami rasti objektai ribojančiais keturkampiais.

Remiantis tikslumo, jautrumo ir F1 metrikomis (aptartos 1.5. skyriuje kartu su IoU objektų

keturkampių sankirtos ir sąjungos santykio efektyvumo verte) įvertinama Otsu binarizavimo metodo efektyvumas mikroplastikų dalelių lokalizavime su skirtingomis tiesinės lygties (4.1.) α ir β reikšmėmis. Eksperimente panaudoti visos tyrimui gautos nuotraukos, apskaičiuojant visų vaizdų vidurkių metrikas, tam, kad rastume optimaliausias šviesumo ir kontrasto (α ir β) reikšmes. Žemiau pateiktame (5 pav.) galime pamatyti 6 grafikus, kur kiekvienas iš jų atspindi metrikų vidurkių priklausomybę nuo α ir β reikšmių. Pirmuose grafikuose matome ryškų jautrumo padidėjimą apytikriai nuo 100 β koeficiento, tai parodo didelį klaidingai teigiamų FP (angl. *False positive*) atpažintų objektų kiekį, o vidutinės tikslumo ir F1 vertės siekia maksimumą ties β 120–140, kai α yra lygi 0.7, 0.8 ir 0.9. Tai rodo, kad mažesnis kontrastas su didesne šviesumo verte veikia efektyviau nei likusieji grafikai, kurie pateikia dar didesnį metrikų kritimą. Taip yra dėl to, kadangi didėjant α ir β vaizdas tampa per šviesus, o tai trukdo algoritmui tiksliau identifikuoti mikroplastiko objektus. Efektyviausi rezultatai fiksuojami ties F1 0.27 metrikos verte, kai $\alpha = 0.8$ ir $\beta = 140$.

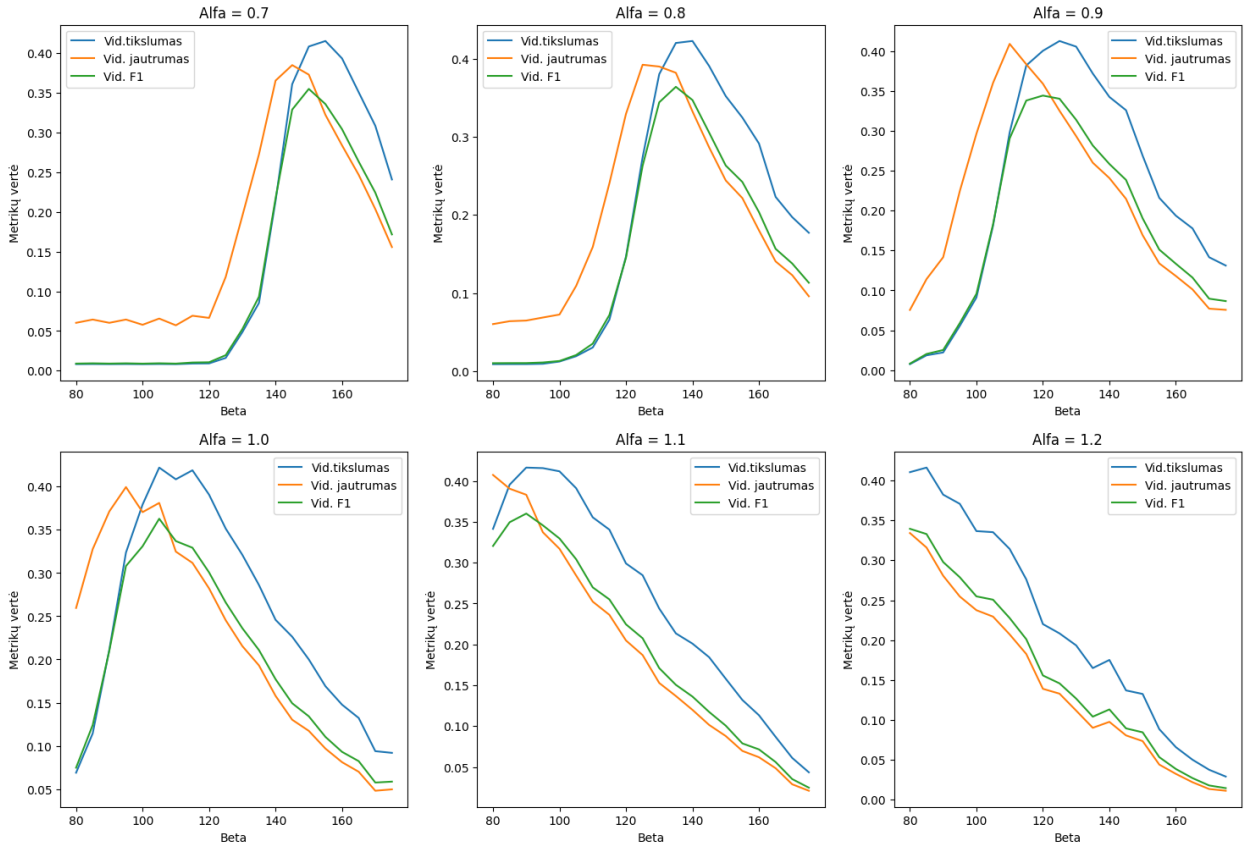


5 pav. Tikslumo, jautrumo ir F1 vidurkių metrikų palyginimas esant skirtingoms α ir β koeficientų reikšmėmis.

Tyrimo eigoje nuspręsta, norint padidinti algoritmo tikslumą būtina atlikti objektus ribojančių stačiakampių filtravimą, tam kad sumažintume klaidingai aptiktų objektų skaičių. Nepakeitus nuotraukos dydžio buvo rasta iš tyrimui skirtų nuotraukų rinkinio mažiausia ribojančio keturkampio ploto reikšmė 161, kurią nustatome kaip minimalią ribą. Tai reiškia, kad aptiktos dalelės, ribojančio keturkampio plotas būdamas mažesnis už nustatytąją minimalią reikšmę yra laikomas netinkamu todėl yra pašalinamas iš tolimesnio tikslumo vertinimo, siekiant sumažinti FP klaidingų teigiamų atvejų skaičių.

Žemiau pavaizduotame (6 pav.) su tokiais pat α ir β koeficientais galime matyti, kaip aptiktų

mikroplastiko dalelių jautrumas ir kiti parametrai pasikeitė palyginant su 5 pav. esančiais grafikais. Sumažėjęs jautrumas padėjo pasiekti optimalesnius rezultatus, nors geriausios metrikos išlieka pagal grafikus kai $\alpha = 0.8$, tačiau tikslumas išauga beveik 20 procentų (0.42), o F1 metrika pasiekė 0.36 vertę ties $\beta = 135$.



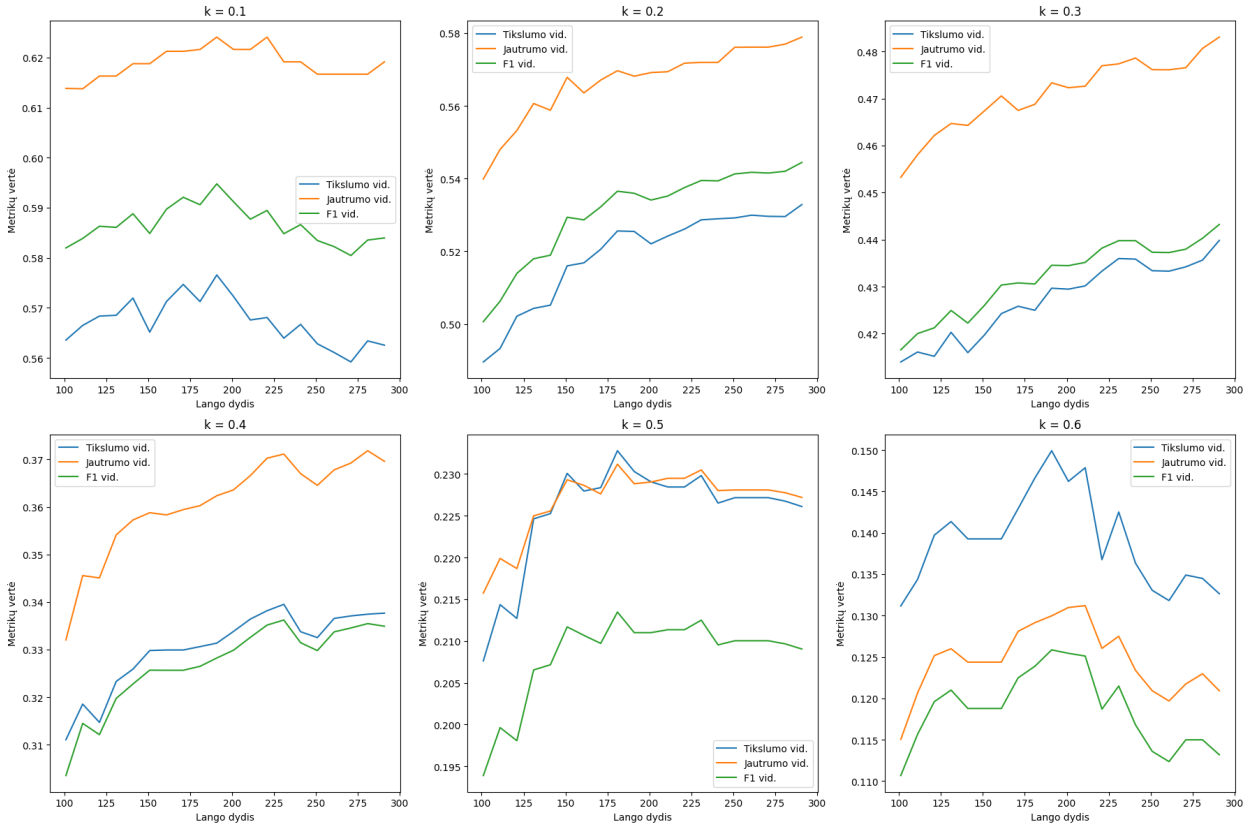
6 pav. Tikslumo, jautrumo ir F1 vidurkių metrikų kitimas pašalinus klaidingai aptiktus objektus pagal ribojančių keturkampių plotą.

4.2.2. Sauvola binarizavimo tyrimas

Priešingai nei Otsu binarizavimo tyrimo etape Sauvola metodas (1.3.2. skyrius) nereikalauja papildomų vaizdo šviesos ir kontrasto modifikavimų, kadangi šis metodas automatizuotai prisitaiko prie skirtingo vaizdo šviesumo naudojant tam tikrus parametrus. Šiam metodui prireiks surasti lango dydį (angl. *window size*) ir k konstantą, kurios parodys su kokiomis jų vertėmis gausime optimaliausius rezultatus mikroplastiko dalelių aptikimo užduotyje. Taip pat dažniausiai tyrimuose naudojama R standartinio nuokrypio reikšmė, kuri parenkama pagal vaizdo pikselių esamą gylį. Šiame tyrime naudojamos pilko fono nuotraukos su 8 bitų gyliu, kurios standartinis nuokrypis dažniausiai moksliniuose tyrimuose [21] būna pasirenkamas pusės viso pikselio intensyvumo diapazono (0 - 255), todėl $R = 128$. Sauvola binarizavimo metodas įgyvendintas Colab platformoje naudojant mašininio mokymosi *scikit-learn* biblioteką [25].

Apačioje pateikti (7 pav.) 6 grafikai, kurie parodo Sauvola metodo vidutines metrikų dydžius, vaizdų binarizavime, priklausomai nuo lango dydžio ir k reikšmės. Tai leidžia analizuoti ir palyginti skirtingų parametrų įtaką algoritmo veikimui turint unikalias mikroplastiko mikroskopijos nuotraukas. Grafikuose galime pamatyti, kad metrikų vertės yra pačios aukščiausios (F1 0.59), kai k intensyvumo koeficientas lygus 0.1, o lango dydis 191, tai leidžia teigti, kad su mažiausia k

verte aptinkama didžiausias tikrųjų mikroplastiko dalelių skaičius, tačiau taip pat ir su išaugančiu klaidingai aptiktų mikroplastiko dalelių skaičiumi. Didėjant k vertei, metrikų reikšmės mažėja, paskutiniame grafike, kur k= 0.6 gauti rezultatai yra prasčiausi, tai rodo, kad kuo didesnė slenkstinio intensyvumo vertė tuo prasčiau binarizavimo metu išskiriamos mikroplastiko daleles šiuose unikaliuose mikroskopijos nuotraukose. Ši tendencija rodo, kad k vertės reguliavimas su langų dydžiais yra svarbus kriterijus Sauvola metode ieškant tinkamiausių verčių aptinkant objektus vaizduose.



7 pav. Tikslumo, jautrumo ir F1 vidurkių metrikų palyginimas esant skirtingoms k ir langų dydžio vertėmis.

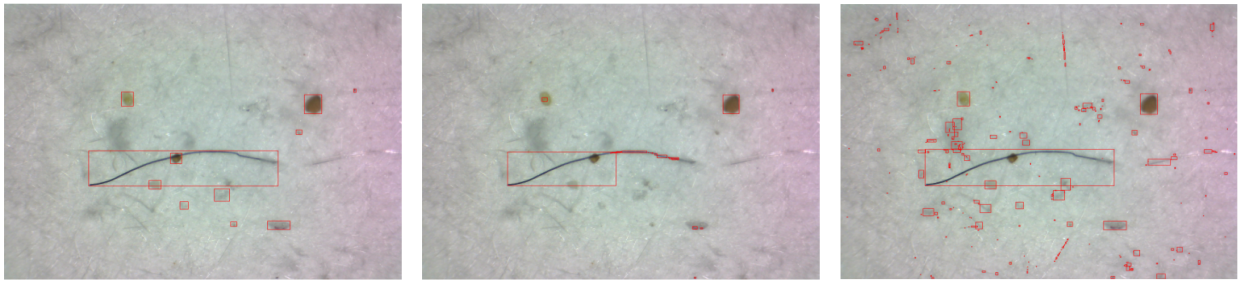
4.3. Otsu ir Sauvola ekperimentų rezultatų apžvalga

Po atliktų eksperimentų, gautos didžiausios vidutinės metrikų vertės rodo, kad naudojant Otsu binarizavimą su parametrais $\alpha = 0.8$, $\beta = 135$ ir kontūrus ribojančių keturkampių plotų filtrą (6 pav.), gauname mažesnį efektyvumo lygį nei su Sauvolos metodo parametrais $k = 0.1$ ir langų dydžiu 191. Tai reiškia kad Sauvola metodas geriau prisitaiko prie didelio kontrasto tarp objektų ir fono rasdamas didesnį kiekį tikrųjų aptiktų mikroplastiko objektų.

Sauvolos tyrimo pradžioje buvo svarstyta atlikti objektų ribojančių kontūrų keturkampių filtravimą Sauvolos binarizavimui, tačiau pastebėta nuotraukų rezultatų pavyzdžiuose, kad toks būdas nėra optimalus, kadangi aptiktuose vaizduose objektai yra suskaidyti į atskirus kontūrus ribojančius keturkampius. Pavyzdžiui, 8 pav. (b) vienalytė gija yra suskaidyta į kelias dalis, o 9 pav. (b) Sauvolos metodu užfiksuotos gijos suskaidymas pateikia kitokį rezultatą nei viename iš tyrimo projektų padarytai binarizacijos išvadai, kad Sauvolos binarizavimo algoritmas padeda išvengti mikroplastiko gijų suskaidymo [17]. Tad naudojant prieš tai aptartą filtrą būtų prarastos mažos gijų, granulių suskaidytos dalys, be kurių morfologinis sprendimas (1.6. skyrius) prarastų savo

veiksnumą.

Toliau pagal žemiau pateiktą 8 pav. palyginsime aptiktų mikroplastiko dalelių nuotraukų rezultatus. Pirmajame eksperimente, Otsu metode (8 pav. (b)), rezultatai parodo, kad netik objektai yra suskaidomi į dalis, bet ir didžioji dalis faktinių objektų nėra aptikti. Sauvolos binarizavimo metodo nuotraukoje (8 pav. (c)) aptikti beveik visi mikroplastikai išskyrus vieną dalelę kuri kontūrais susijungusi su gija, tačiau toks atvejis šiame duomenų rinkinyje yra vienetinis, todėl vykdant tolimesnius šio projekto eksperimentus nebus bandoma spręsti šios problemos. Tačiau pasirinkti Sauvolos algoritmo parametrai (k ir lango dydis) šiuo tyrimo atveju turi didelį kiekį aptiktų neegzistuojančių objektų.



(a) Vaizdas su tikromis mikroplastiko lokacijomis.

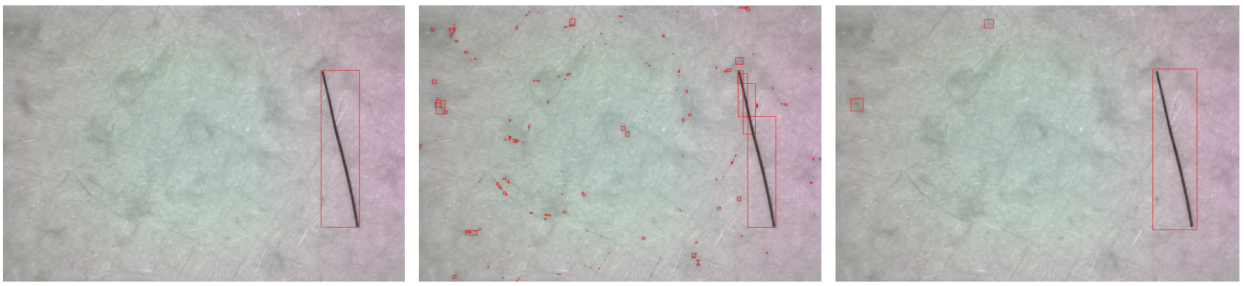
(b) Otsu, aptikti nevisi faktiniai objektai, gija suskaidyta į keletą dalių.

(c) Sauvola, užfiksuoti objektai su dideliu klaidingai aptiktų objektų skaičiumi.

8 pav. Binarizavimo metodų palyginimas objektų aptikime nenaudojant ribojančių keturkampių ploto filtro.

Naudodami Otsu metodą gauname aukščiausią 0.2 tikslumo (angl. *precision*) ribą objektų aptikime, o Sauvola metodas aptinka didesnį kiekį (0.56 tikslumas) faktinių mikroplastiko dalių, todėl tolimesnei projekto eigai skirsime dėmesį šiam efektyvesniam metodui. Nors didesnis faktinių objektų radimas Sauvola metode lemia ir didesnį klaidingai aptiktų objektų skaičių, bei tokią pačią problemą kaip ir Otsu metode - objektų skaidymas į atskiras dalis, naudodami morfologines operacijas galime sumažinti šių problemų tikimybę suderindami su plėtimosi, erozijos, atidarymo ir uždarymo, mažų objektų šalinimo operacijos metodais. Morfologinės operacijos atliktos *Google Colab* platformoje pasitelkus *OpenCV* bibliotekomis.

Žemiau pateiktuose nuotraukose (9 pav. (b)) galime pamatyti juodą giją, kuri yra suskaidyta į kelias dalis ir kitus objektus kurios yra traktuojamos kaip klaidingai aptiktos mikroplastiko dalelės. Pati gija nėra išskaidyta dideliais tarpais, dėl to pasirinktas 5x5 stačiakampis, struktūrinis elementas ir panaudota atidarymo operacija, kuri užpildė nedidelius baltų pikselių tarpus juoda spalva taip sujungdami objektą į vieną giją. Tačiau po šios morfologinės apdorojimo procedūros vis tiek išliko didelis kiekis aptiktų ne mikroplastiko objektų, todėl buvo įgyvendinti papildomi morfologiniai žingsniai: plėtimasis, erozija, uždarymas. Atlikę keletą eksperimentinių bandymų koreguojant elemento branduolio dydį ir morfologinių operacijų iteracijų skaičių kiekį buvo atrasti optimaliausi parametrai keliuose nuotraukų rinkiniuose. Lango branduolio dydis išliko toks pats, o pirmoji panaudota plėtimosi funkcija vykdoma 4 kartus tol kol pašalinome kuo didesnį kiekį netiksliai aptiktų mikroplastiko objektų, tuomet vykdomas erozijos 14 kartų iteracija, kurios sprendimas šiuo atveju yra užpildyti gijos pašalintus juodus pikselius, kadangi po plėtimosi operacijos ši linijinė gija išskaidoma. Vaizdas (9 pav. (b)) yra gerokai patobulintas (9 pav. (c)), o mikroplastiko gija po apdorojimo liko išsaugota ir didžioji dalis nereikalingų objektų - sėkmingai pašalinti.



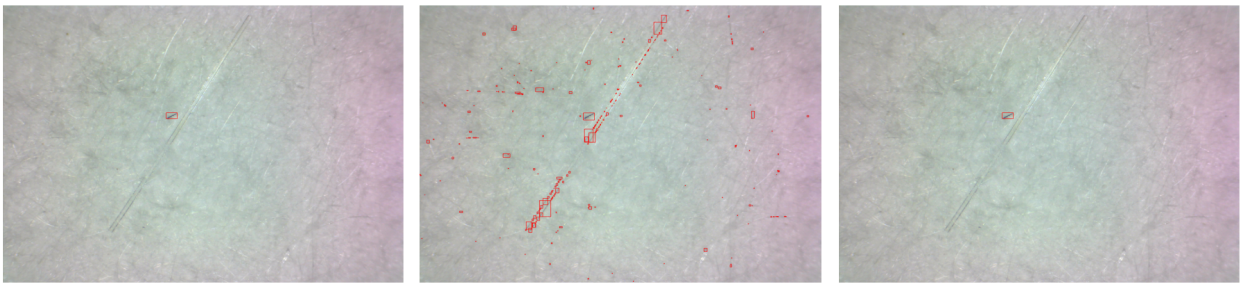
(a) Vaizdas su tikra mikroplastiko lokacija.

(b) Vaizdas po Sauvolos binarizacijos su klaidingai suskaidyta gija.

(c) Vaizdas po morfologinio apdorojimo.

9 pav. Sauvola binarizavimo metodo dalelių išsiskyrimas ir morfologijos įtaka vaizdui.

Kitame pateiktame pavyzdyje (10 pav.) dažnai pasitaikančioje problemoje po adaptyviosios binarizacijos metodo taikymo, galime pamatyti klaidingai aptiktas pilkas mikroskopo fiksuotas filtro dėmes, kurios yra visame vaizdų duomenų rinkinyje. Pasitelkus morfologinį sprendimą su prieš tai minėtomis funkcijomis ir iteracijų skaičiumi buvo pasiektas optimalus rezultatas (10 pav. (c)), kadangi pavyko pašalinti netinkamus objektus, o tikrasis mikroplastiko objektas išliko pažymėtas kontūru ribojančiu keturkampiu kurio slenkstinė vertė visame darbe nekito 0.5.



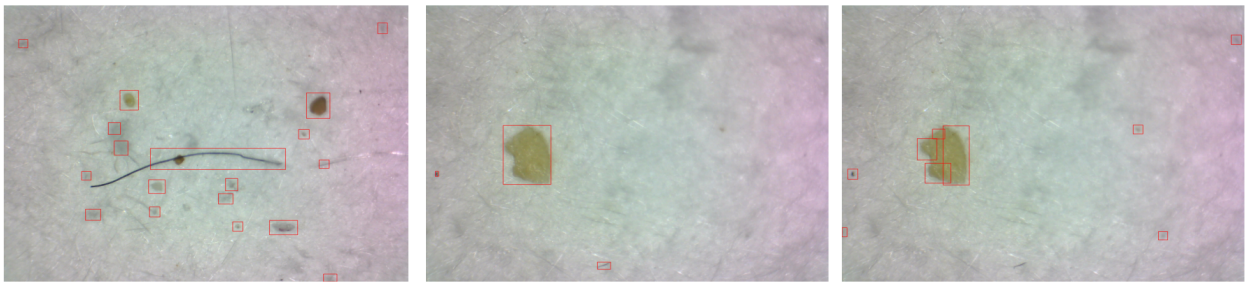
(a) Vaizdas su tikra mikroplastiko lokacija.

(b) Užfiksuotos pilkos mikroskopo filtro dėmės.

(c) Vaizdas po morfologinio apdorojimo.

10 pav. Sauvolos binarizavimo metodas, pilkų mikroskopijos filtro dėmių išsiskyrimas ir morfologijos įtaka vaizdui.

Tačiau prieš tai analizuoti morfologinių parametrų rodikliai kitokiame duomenų išskirtame rinkinyje nėra tinkami, kadangi pažvelgus žemiau į 11 pav. galime palyginti 11 pav. (a) su 8 pav. (c) gautus vaizdų rezultatus. Centrinė mikroplastiko gija praradusi dalį kontūro ribų, o fragmentų, granulių ribos padidėjusios neatitinka faktinių objektų formų. Šis pasirinktas morfologinis plėtimosi ir erozijos sprendimas neišsprendė visų mikroplastiko skaidymo problemų, jeigu toliau didintume morfologijos parametrus prarastume dar daugiau aptiktų TP faktinių objektų, kurie sukeltų klasifikavimo modelio persimokymą dėl ribotų duomenų kiekio. Tai reiškia, kad modelis būtu nepajėgus efektyviai atlikti mokymo sprendimus su naujais, anksčiau nematytais duomenimis.



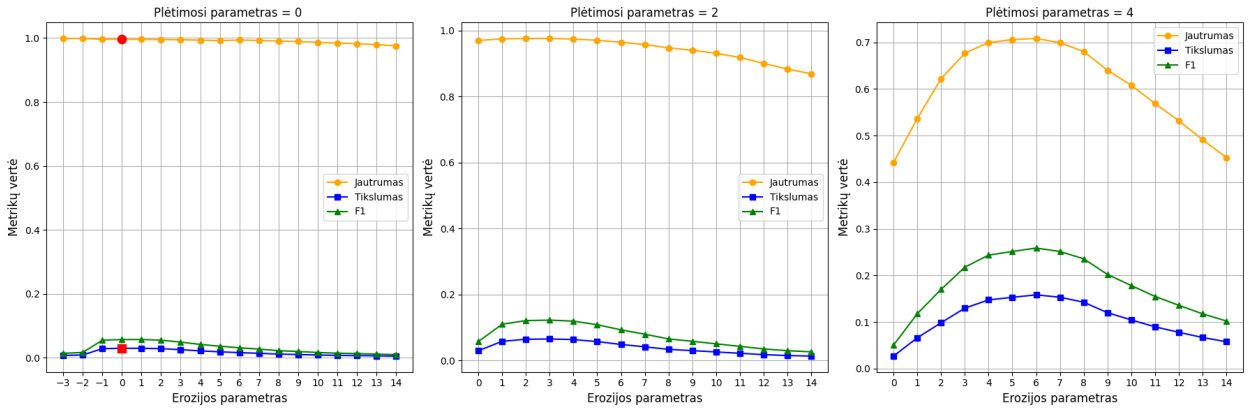
(a) Objektai praradę savo faktinę formą.

(b) Vaizdas su tikra mikroplastiko lokacija.

(c) Panaudotas vaizdo morfologinis apdorojimas, granulė nėra visiškai ištaisyta.

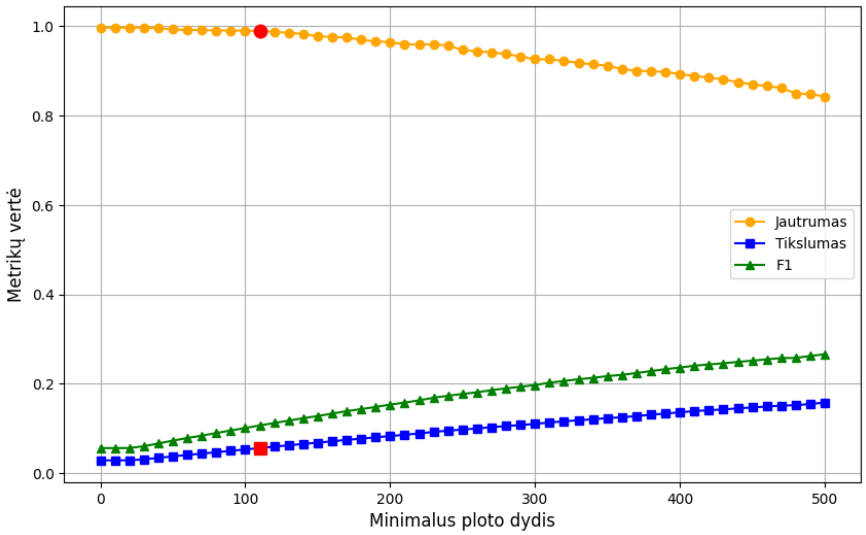
11 pav. Sauvola binarizavimo ir morfologijos taikymo problemos mikroplastiko aptikime.

Žemiau pateiktuose grafikuose (12 pav.) parodoma kaip kinta nuotraukų jautrumo ir tikslumo vertinimo metrikos nenaudojant morfologinio sprendimo ($x = -3$), naudojant tik atidarymo ($x = -2$), tik uždarymo ($x = -1$), atidarymo ir uždarymo parametrus ($x = 0$) bei erozijos reikšmių diapazonu nuo 0 iki 14 pritaikant plėtimosi parametrus 0, 2 ir 4. Didžiausias jautrumas (kuo didesnė vertė tuo daugiau teisingai aptiktų objektų) yra pirmame taške (-3: $TP = 763$, $FP = 112488$, $FN = 1$), kai nėra naudojami jokie morfologiniai procesai, o antrame taške (-2: $TP = 768$, $FP = 93266$, $FN = 1$) gaunami nežymiai didesni tikslumo metrikos rezultatai (tikslios TP, FP, FN vertės išsaugojamos csv formato faile), kadangi naudojant tik morfologinį atidarymo procesą aptinka 5 vienetais daugiau tikrų mikroplastiko dalelių, o FP tik truputį nusileidžia taškui -3. Taikant iškarto abu atidarymo ir uždarymo procesus gauname geriausią santykį tarp visų gautų tiksliai ir klaidingai aptiktų objektų rezultatų (plėtimosi parametras 0, 0: $TP = 763$, $FP = 25480$, $FN = 3$). Sumažėjus FP matome grafike tikslumo metrikos augimą, tai reiškia, kad mikroplastiko dalelių susiskaidymas sudarė didžiąją dalį problemos Sauvolos binarizavimo metode šiam vaizdų duomenų rinkiniui. Toliau pasitelkus erozijos ir didesnius plėtimosi parametrus pavyksta sumažinti klaidingai aptiktų faktinių objektų skaičių tačiau, kaip ir prieš tai aptarėme kuo didesni morfologinės operacijos parametrai tuo tikrieji mikroplastikai pernelyg deformuojami. Dėl to gijų, granulių ir likusių mikroplastiko fragmentų formos gali supanašėti, tuomet klasifikatoriui būtų sudėtinga išskirti pagrindines kiekvieno objekto savybes. Taip pat plėtimasis ir erozija mažina tinkamai aptiktų objektų kiekį bei atsiranda vis daugiau FN visiškai neaptiktų mikroplastiko dalelių, dėl šių priežasčių toliau darbo projekte bus taikoma tik atidarymo ir uždarymo morfologiniai procesai.



12 pav. Tikslumo, jautrumo ir F1 metrikų kitimas pritaikius morfologinius sprendimus.

Dar viename eksperimente buvo siekiama padidinti lokalaus binarizavimo tikslumą sumažinant klaidingai tikrų objektų aptiktų atvejų skaičių FP, pernelyg nesumažinant jautrumo metrikos. Pritaikius mažų objektų šalinimo morfologinę operaciją po atidarymo ir uždarymo proceso buvo gauti optimalesni rezultatai (13 pav.), didinant minimalaus šalinimo ploto diapazoną kas 10 ties 100 verte galime pamatyti aiškesnį tikslumo augimą bei jautrumo mažėjimą. Detaliau pažvelgus į gautus rezultatus, $TP = 763$ liko nepakitęs iki 120 šalinimo ploto vertės, FP pasiekė vos ne pusę savo buvusios reikšmės ($FP = 12727$) kai buvo naudojama tik atidarymo uždarymo operacija, o FN visiškai neaptiktų tikrų objektų skaičius padidėjo 6 vienetais ($FN = 9$), kai buvo tik 3. Tolimesni grafiko rezultatai rodo, kad šis mažų objektų šalinimo metodas pasiteisino ir iš tiesų sumažino klaidingai aptiktų mažų objektų skaičių, likusios grafiko vertės didesnės nei 110 pateikia aukštesnį objektų aptikimo tikslumą tačiau tuo pačių prarandamos tikrųjų mikroplastiko dalelių kiekis.



13 pav. Tikslumo, jautrumo ir F1 metrikų kitimas pritaikius morfologijos mažų objektų šalinimo metodą.

Iš šių atliktų analizių galime teigti, kad Sauvola binarizavimo metodas yra efektyvesnis nei Otsu metodas, kadangi aptinka didesnį tikrų teisingų objektų kiekį, kurie yra svarbūs tolimesnei klasifikavimo užduočiai spręsti. Sauvola metodas, nors ir yra pranašesnis taikant šiame mikroskopijos vaizdų rinkinyje, taip pat susidūrė su problemomis, kaip ir Otsu metodas, tokiais kaip dalelių skaidymas, kas padidina TP atvejų kiekį, taip pat, kaip mikroskopijos filtro dėmių fiksavimas. Tačiau, morfologinių veiksnių pritaikymas padėjo sumažinti TP skaičių, leidžiant gauti tikslesnį metrikų įvertinimą ir pagerinti bendrą modelio veikimą. Būsimuose tyrimuose bus naudojami atidarymo, uždarymo ir minimalaus ploto šalinimo (110 plotas) morfologiniai procesai, kurie optimizuoja klaidingai teigiamų atvejų mažinimą, ne mažinant tikrų teigiamų objektų atvejų skaičiaus tam, kad klasifikatoriaus modelio veikimas kuo tiksliau atspindėtų realių sąlygų efektyvumą.

5. Mašininio mokymo klasifikatorių taikymas

Šiame skyriuje taikomi aptarti (2. skyrius) mašininio mokymosi klasifikavimo metodai, kurių eksperimentai vykdomi taikant dviejų klasių klasifikacijai (angl. *Binary classification*). Aptiktų objektų vaizdų duomenų rinkiniui naudinga kuomet siekiama klasifikuoti duomenis tik į dvi klases, šio projekto atveju pirmoji klasė - mikroplastikas, antroji klasė - ne mikroplastikas. Toks metodo taikymas aiškiai konkretizuoja pagrindinę šio projekto atliktų Sauvolos ir Otsu eksperimentų užduotis - kuo tikslesnis mikroplastiko aptikimas, kadangi toks modelio panaudojimas pasitelkus specifines mikroplastiko charakteristikas (5.1. skyrius) padeda sumažinti klaidingai teigiamų ir klaidingai neigiamų atvejų skaičių, padidinant bendrą teisingai atpažintų objektų tikslumą. Kelių klasių klasifikacijos metodas suteikia galimybę atlikti gilesnę analizę identifikuojant įvairias mikroplastiko formas.

Mašininio mokymosi procese aptiktų objektų duomenų rinkinys yra padalintas į tris skirtingo dydžio dalis: mokymo (angl. *training*), testavimo (angl. *test*) ir validavimo (angl. *validation*) rinkinius. Mokymo dalį sudaro 60 procentų viso rinkinio duomenų, likusieji pasidalina po 20 procentų. Didžiausias duomenų rinkinys yra skirtas klasifikatoriaus modelio mokymui, kuris pagal duomenų ryšius, charakteristikas išmoksta prognozuoti duomenis, o validavimo rinkinys padeda nustatyti mokymo proceso metu, kaip gerai modelis veikia su duomenimis, kurių nematė pačio mokymosi metu, tai leidžia išvengti modelio persimokymo bei parinkti optimalesnius klasifikatoriaus parametrus. Testavimo rinkinys naudojamas jau galutiniam apmokyto modelio efektyvumui ir tikslumui vertinti, šiuos duomenis modelis pamato pirmą kartą, kadangi svarbu patikrinti, kaip modelis veikia su naujais, nematytais duomenimis.

Visuose klasifikavimo medeliuose naudojamas stratifikuotas kryžminis patvirtinimas (angl. *Stratified cross validation*), kuris užtikrina kiekvienos klasės duomenų proporcingą padalijimą į skirtingus mokymo (modelio mokymui), testavimo (galutinio modelio įvertinimui) ir validavimo (geriausių parametrų radimui) rinkinius. Tai sumažina riziką, kad modelis bus netinkamai apmokytas dėl nevienodų klasių padalijimo, kurį gali sukelti naudojant paprastą kryžminį padalijimą (angl. *Cross validation*). Dažniausiai yra renkamasis 10 kartų kryžminė patikros padalijimas, nes tai sudaro gerą balansą tarp tikslumo ir skaičiavimo trukmės (priklausomai nuo turimų duomenų ir parinktų parametro tinklelio derinimo variantų) [1].

5.1. Objektų savybių išskyrimas ir jų etikečių nustatymas

Aptiktų objektų savybių išskyrimo etapas pasitelktus Suzuki ir Abe metodą (1.4. skyrius) yra svarbus etapas tolimesnei mikroplastiko aptikimo tikslumo efektyvumui gerinti, užduočiai atlikti. Iš visų esamų faktinių anotuotų mikroplastiko kategorijų (fragmentai, gijos, granulės) pagrindinės išgaunamos charakteristikos suskirstomos į spalvines ir geometrines savybes, kurių gauti unikalūs ir specifiniai mikroplastiko kategorijų duomenys yra esminiai norint efektyviai identifikuoti ir klasifikuoti tikruosius mikroplastiko objektus. Išnagrinėjus turimą vaizdų duomenų rinkinį kiekviena mikroplastiko kategorija turi sau būdingas savybes, pavyzdžiui gijos yra ilgos ir plonos dažniausiai jų perimetras yra didesnis nei kitų mikroplastiko rūšių, dažniausiai vaizdų duomenų rinkinyje aptinkami fragmentai išsiskiria mažu dydžiu su įvairiomis formomis, o granulės įprastai yra ne juodos spalvos, kaip kitos kategorijos, jų išskirtinė savybė - formos ovalumas, neturi smailių kampų ar išsikišimų [17]. Žemiau pateikta 1 lentelė su išgautomis geometrinėmis savybėmis:

1 lentelė. Aptiktų objektų išskirtos geometrinės atributų etiketės su aprašymu.

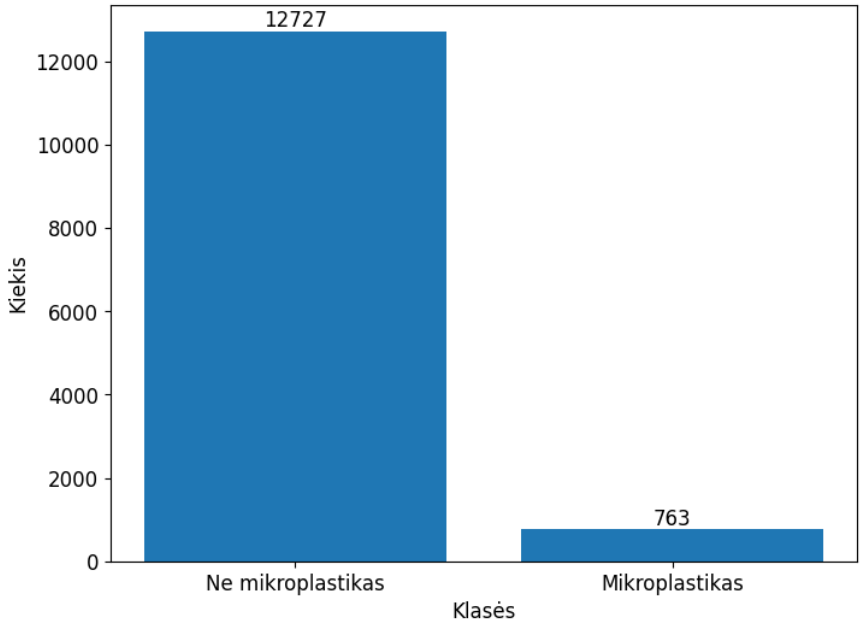
Etiketė	Aprašymas
area	Plotas pikseliais nurodantis objekto dydį.
perimeter	Perimetras pikseliais nurodantis objekto ilgumą.
form_factor	Formos faktorius nurodantis objekto apvalumą, kuo skaičius artimesnis 1, tuo objektas yra panašesnis į apskritimą.
compactness	Kompaktiškumas nurodo ne tik apie objekto formos apvalumą, bet ir yra naudingas vertinant objekto išsiplėtimą, lyginant perimetrą su plotu.
blob_to_bbox_ratio	Objekto ploto ir kontūro ribojančio keturkampio santykis, kuo skaičius artimesnis 1, tuo objektas užima didesnę dalį ribojančio stačiakampio, tai labiau būdinga fragmentų ir granulių mikroplastikui.
bbox_ratio	Objekto pločio ir aukščio santykis su kontūro ribojančiu keturkampiu, kuris nurodo, kaip stipriai objekto esama proporcija yra nutolusi į plotį ar aukštį, toks atributas naudingas ieškant ilgų gijų, stambių fragmentų arba granulių.
radius_ratio	Atstumas nuo tolimiausio ir artimiausio kontūro centrinio pikselio, kuris nurodo kiek arti centro yra objekto kontūras, tai svarbu nustatant, ar objektas turi įdubimus, išsikišimus.

Spalvines savybes sudaro RGB spalvų diapazonas ir atspalvio, prisotinimo ir vertės (angl. *Hue, Saturation, Value, HSV*) spalvų apibūdinimo modelis. Šis modelis nurodo objekto dominuojančią spalvą, įvertindama kaip ryškiai yra prisotinta ta dominuojanti spalva ir vertė nurodanti spalvos šviesumo, tamsumo lygį. Žemiau pateikta 2 lentelė su visomis išskirtomis spalvinėmis savybėmis:

2 lentelė. Aptiktų objektų išskirtos spalvinės atributų etiketės su aprašymu.

Etiketė	Aprašymas
gray_avg	Pilkos skalės vidutinis šviesumo lygis nurodantis viso objekto bendrą spalvos intensyvumą.
gray_var	Pilkos skalės šviesos kintamumo (variacijos) rodiklis nurodantis kaip stipriai spalvos lygis kinta objekte.
r_avg	Visų, raudonos spalvos diapazono (RGB modelis), pikselių vidurkio vertė.
r_var	Visų, raudonos spalvos diapazono (RGB modelis), pikselių variacijos vertė.
g_avg	Visų, žalios spalvos diapazono (RGB modelis), pikselių vidurkio vertė.
g_var	Visų, žalios spalvos diapazono (RGB modelis), pikselių variacijos vertė.
b_avg	Visų, mėlynos spalvos diapazono (RGB modelis), pikselių vidurkio vertė.
b_var	Visų, mėlynos spalvos diapazono (RGB modelis), pikselių variacijos.
h_avg	Dominuojančios spalvos vidutinė pikselių vertė (HSV modelis).
h_var	Dominuojančios spalvos variacijos pikselių vertė (HSV modelis).
s_avg	Dominuojančios spalvos vidutinis ryškumo įvertis (HSV modelis).
s_var	Dominuojančios spalvos variacijos ryškumo įvertis (HSV modelis).
v_avg	Vidutinis šviesumo lygis nurodantis viso objekto bendrą šviesumo, tamsumo intensyvumą.
v_var	Variacijos šviesumo lygis nurodantis viso objekto bendrą šviesumo, tamsumo intensyvumą.

Sudarius šiuos atributus toliau nustatomos dviejų klasių etiketės: ne mikroplastikas ir mikroplastikas. Šios etiketės prie kiekvieno aptiktų Sauvolos binarizacijos ir pritaikytų morfologinių reikšmių objektų sudarė jau žinomas TP ir FP objektų kiekių rezultatus, kurios pavaizduotos apačioje 14 pav. (visos išskirtos objektų savybės kartu su jų rastu klasių etiketėmis eksperimento metu įtraukiamos į *csv* formato failą):



14 pav. Mikroplastiko klasės ir jų pasiskirstymas.

5.2. Atsitiktinis miškas

Šio skyriaus apačioje, pateiktoje 4 lentelėje yra užfiksuoti taikyti skirtingi atsitiktinio miško modelio eksperimentų bandymų rezultatai, kurie buvo atlikti siekiant išnagrinėti modelio efektyvumą įvairiose parametrų deriniuose ir papildomų metodų taikymuose. Pirmasis atsitiktinio miško apmokymas vykdomas naudojant visus 21 išskirtus atributus (5.1. skyrius). Pritaikant parametrų tinklelio sprendimą ieškant efektyviausių klasifikatoriaus modelio rezultatų, kurie apžvelgti 2.1. skyriuje, jų atsitiktinai pasirinktos reikšmės pateiktos 3 lentelėje. Toliau visi lentelėje nurodyti parametrai yra taikomi nekartu ir nepriklausomai nuo jų dabartinės išdėstymo tvarkos.

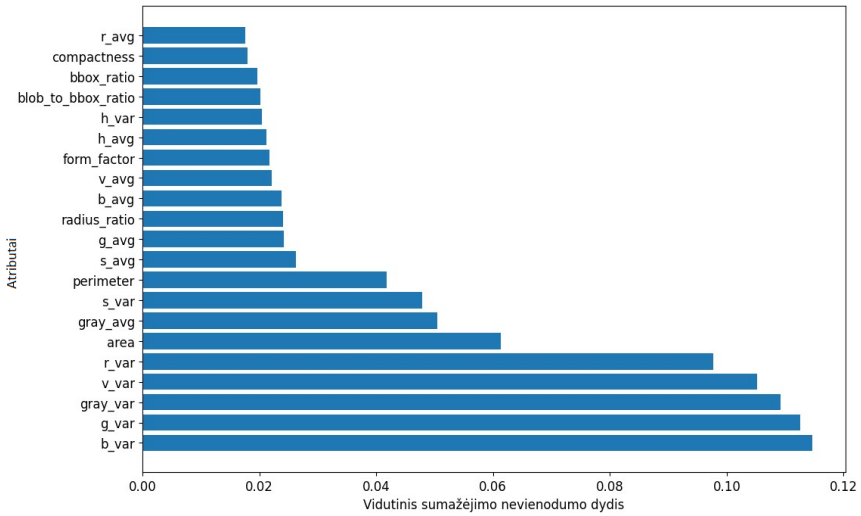
3 lentelė. Atsitiktinio miško klasifikatoriuje taikytos parametrų tinklelio reikšmės.

Parametrai	Reikšmės
n_estimators	10, 25, 50, 100, 200, 300, 400
max_features	auto, sqrt, log2
max_samples	None, 0.1, 0.2, 0.3, 0.4, 0.5
max_depth	None, 10, 20, 30, 40, 50
min_samples_split	2, 5, 10, 20, 30
min_samples_leaf	1, 2, 4, 10
bootstrap	True, False
criterion	gini, entropy

Tinklelis nustatomas atsižvelgiant į *scikit-learn* bibliotekos esamus parametrų pavadinimus:

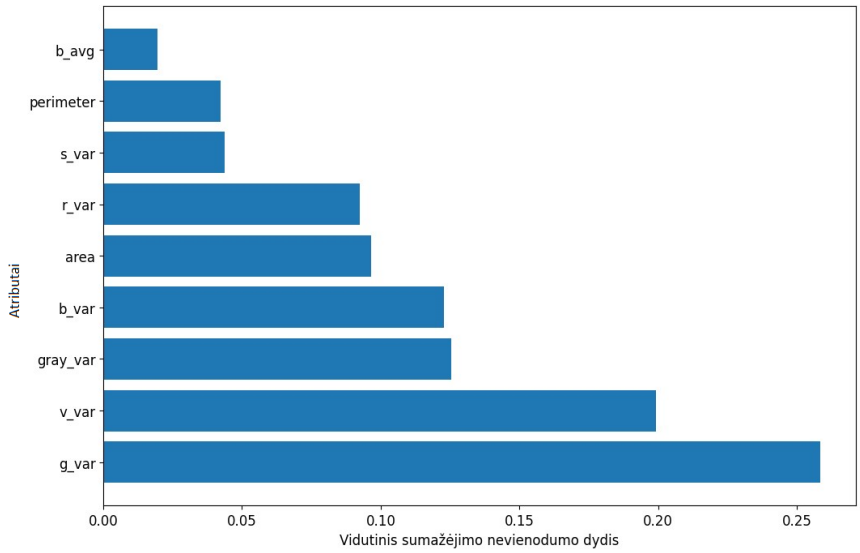
medžių skaičius (angl. *n_estimators*), atsitiktinai parenkamų kintamųjų kiekį (angl. *max_features*), mėginių dydis (angl. *max_samples*), pakartojimas ir skaidymo taisyklė (angl. *criterion*). Taikant tik prieš štai paminėtus parametrus modelis sukuria 200 medžių, kurio kiekvienam medžiui naudojama skirtingą (su pakartojimu) 40 procentų mėginių dydį (duomenų kiekis iš viso rinkinio), o jų mokymui atsitiktinis kandidatų skaičius parenkamas iš visų tame rinkinyje esančių savybių. Šis apmokytas modelis naudojantis nesubalansuotomis klasių dydžiais (14 pav.) ir entropijos skaidymo taisykle pasiekė, kad iš visų prognozuotų teigiamų mikroplastiko objektų 0.67 atvejų buvo teisingai atpažinti, o jautrumas iš visų teigiamų mikroplastiko objektų, kuriuos modelis turėjo atpažinti, buvo teisingai identifikuoti 59 %. Kitas parametrų būdas atliekamas pašalinant iš parametrų tinklelio mėginių dydžių parametrus, tai modeliui leidžia naudoti kiekvienam medžio mokymui visus duomenų rinkinio duomenis taikant galimai tų pačių savybių mokymo pakartojimą. Gauti rezultatai prastesni: tikslumas 0.68, jautrumas 0.58, F1 0.62. Kitas taikymas vyko nenaudojant duomenų pakartojimo dėl kurio rezultatai tapo dar prastesni: tikslumas 0.66, jautrumas 0.57, F1 0.61. Tai reiškia, kad duomenų pakartojimo naudojimas modelio mokymo procese turi reikšmingą poveikį modelio našumui. Sekančiame bandyme pritaikomi papildomi *scikit-learn* bibliotekos siūlomi atsitiktinio miško parametrai: didžiausias medžių gylis, mažiausias skaidymui reikalingų mėginių skaičius ir mažiausias galutinio mazgo turimų mėginių skaičius. Šios naujai pridėtos atsitiktinio miško taisyklės prie parametrų tinklelio padėjo modeliui pasiekti nežymiai aukštesnius rezultatus (4 lentelė, Parametrų derinimas): tikslumas 0.69, jautrumas 0.59 ir F1 0.64. Geriausi modeliui atrinkti parametrai naudoja entropijos skaidymo taisyklę, kai medžių maksimalus gylis 10, o duomenų rinkinio dydis (su pasikartojamu) - 10 procentų. Modelis sukuria 100 medžių ir jų minimalus savybių kiekio skaidymas nebevykdomas kai medyje lieka 2 ar mažiau savybių bei užtikrinant kad po skaidymo medžio mazge liks nemažiau 1 atributo vertė (parametrai: *bootstrap True*, *criterion entropy*, *max_depth* 10, *max_features auto*, *max_samples* 0.1, *min_samples_leaf* 1, *min_samples_split* 2 ir *n_estimators* 100).

Sekančiame klasifikatoriaus mokymo tyrimo etape atliekama atsitiktinio miško modeliui tinkamiausių atributų (5.1. skyrius) paieška, pasitelkus *scikit-learn* bibliotekos MDI vidutinį sumažėjimo nevienodumo skaičiavimą (2.1.1. skyrius). Apskaičiuotos visos savybių MDI atvaizduotos grafike (15 pav.), pagal kurį toliau siekiant išgauti aukštesnius rezultatus galime išskirti 9 (nuo perimetro iki mėlynos spalvos diapazono variacijos intensyvumo parametro (angl. *b_var*) didesnių rodiklių savybes, kurios įtraukiamos į naują modelio mokymą.



15 pav. MDI dydis su visomis mikroplastiko objektų savybėmis.

Gauti rezultatai nesutapo su parametrų derinimo gautais rezultatais, šį kartą skaidymas efektyviausias su visais mėginiais, kurio gylis 10 taikant 200 medžių skaidymą ir jo minimalus skaidymo skaičius 30 su galutiniu mazgo limitu 10. Gauti rezultatai pateikia šiek tiek aukštesnę jautrumo vertę 0.67 nei pirmieji du bandymai, o tai reiškia, kad apmokytas modelis yra efektyvesnis atpažįstant tikruosius teigiamus atvejus, tačiau tikslumo vertė 0.67 nepagerėjo - neigiami atvejai priskiriami į teigiamus, o balanso F1 vertė išliko vienoda 0.64, (4 lentelė, Atributų svarbumo įvertinimas). Geriausi parametrai: *bootstrap True*, *criterion entropy*, *max_depth* 10, *max_features auto*, *min_samples_leaf* 10, *min_samples_split* 30 ir *n_estimators* 200. Nepavykus pasiekti aukštesnių tikslumo rezultatų vykdomas dar vienas modelio vidutinis nevienodumo sumažėjimo verčių skaičiavimas (16 pav.). Pats MDI verčių svarbumas modeliui išaugo iki maksimalios 0.25 vertės, modelio apmokymui naudoti pirmieji trys atributai, kurių vertės viršijo 0.15 MDI dydį, tai - pilko fono spalvų variacija *gray_var*, HSV šviesumo stiprumo variacija *v_var* ir žalios spalvos diapazono variacijos *g_var*.



16 pav. MDI dydis su 2 eksperimento naudotomis mikroplastiko objektų savybėmis.

Palyginus 15 pav. ir 16 pav. galime pamatyti, kad mėlynos spalvos diapazono variacijos vertės prarado svarbiausio atributo poziciją, kai buvo atnaujintas MDI skaičiavimas su trečiame eksperimente naudojamais geriausiais parametrais. Tačiau sumažinus savybių skaičių gauti rezultatai suprastėjo: tikslumas 0.66, jautrumas 0.52 ir F1 0.58. Tai reiškia, kad atributų skaičių mažinimas nėra naudingas, nes modelis prarado svarbius duomenų ryšius.

Svarbu nepamiršti tai, kad šie visi eksperimentai taikomi naudojant nesubalansuotas duomenų klases, todėl siekiant pagerinti klasifikavimo modelio efektyvumą nuspręsta taikyti HNM. Atliktame eksperimente vykdomas 100 iteracijų modelio mokymo procesas, kurio metu stebima kaip modelis susidoroja su dideliu mokymo kartojimo kiekiu teigiamoje klasėje. Modelių rezultatai ganėtinai stabilūs, su pastoviais metrikų svyravimais, kur F1 svyruoja tarp 0.63 iki 0.67. Tai rodo, kad modelis išlaiko optimalų atpažinimo gebėjimą ne per mokant modelio, nepaisant pasirinkto iteracijų kiekio. Aukščiausią tikslumo ir tuo pačiu balanso vertę tarp visų mokymo iteracijų sudarė šie gauti modelio rezultatai: tikslumas 0.79, jautrumas 0.58 ir F1 0.67 (4 lentelė, HNM).

Nors modelio tikslumas taikant sunkių neigiamų pavyzdžių metodą pateikė efektyvesnius rezultatus nei kiti eksperimentai yra svarbu atlikti kitą bandymą, tai - duomenų sumažinimą (angl. *undersampling*). Šis metodas yra naudojamas siekiant išlyginti klasių balansą, atsitiktinai maži-

nant didesnės klasės duomenis. Toks metodas sukuria sąlygas, kur kiekviena klasė turi vienodą svorį mokymo proceso metu. Tai leidžia modeliui išmokti atpažinti kiekvienos klasės ypatybes be papildomo šališkumo, kuris gali atsirasti dėl vienos klasės dominavimo. Gauti geriausi parametrai išliko nepakitę tačiau mikroplastiko klasės tikslumas ženkliai suprastėjo dėl prarastų duomenų ryšių: tikslumas 0.41, jautrumas 0.93 ir F1 0.57 (4 lentelė, Klasių balansavimas). Efektyviausi parametrai: *bootstrap True*, *criterion entropy*, *max_depth* 20, *max_features auto*, *min_samples_leaf* 2, *min_samples_split* 2 ir *n_estimators* 200. Kartu taikant duomenų mažinimą ir SMOTE (2.6. skyrius) metodus buvo pasiekti abiejų klasių 3818 duomenų dydžiai, bet dėja toks mėginimas neefektyvus, rezultatai teigiamai klasei dar labiau suprastėjo: tikslumas 0.38, jautrumas 0.90 ir F1 0.54. Taikant skirtingus balansų dydžius naudojant tik SMOTE, rezultatai išliko panašūs, jautrumas artimas vienetui, o tikslumas žemesnis nei 0.5 vertė.

Remiantis šiame atsitiktinio miško modelio parametrų derinimo eksperimentų rezultatais galime teigti, kad įtraukti nauji parametrai (maksimalus medžių gylis, skaidymo ir galutinio mazgo mėginių limito vertės) minimaliai padėjo padidinti modelio efektyvumą. Antrasis atributų svarbos vertinimo eksperimentas, parodė, kad tinkamas savybių derinys turi įtakos modelio sugebėjimui tiksliau atpažinti teigiamas klases. Trečiasis eksperimentas taikant sunkiųjų neigiamų pavyzdžių metodą klasių nesubalansuotiems duomenis galutiniai rezultatai parodė, kad modeliui pavyko efektyviai mažinti klaidingai teigiamų atvejų skaičių ir padidinti tikslumo metriką be persimokymo pavojaus. Paskutiniame bandyme neigiamos klasės duomenų sumažinimas ir SMOTE dirbtinis klasių balansavimas teigiamoje klasėje drastiškai sumažino modelio gebėjimą tiksliai atpažinti mikroplastiką.

4 lentelė. Atsitiktinio miško klasifikatoriaus įvertinimas atliekant skirtingus taikymus, pagal validavimo duomenų rinkinį.

Eksperimentas	Tikslumas	Jautrumas	F1
Parametrų derinimas	0.69	0.59	0.64
Atributų svarbumo įvertinimas	0.67	0.61	0.64
HNM	**0.79**	**0.58**	**0.67**
Klasių balansavimas	0.41	0.93	0.57

5.3. KNN

Šiame poskyryje K - artimiausio kaimyno klasifikatoriaus modeliui taikyti skirtingi eksperimentų bandymai, kurių rezultatai pateikti 6 lentelėje. Žemiau pateiktoje 5 lentelėje yra nurodytos šiame KNN tyrime taikytos parametro tinklelio reikšmės.

5 lentelė. KNN klasifikatoriuje taikytos parametrų tinklelio reikšmės.

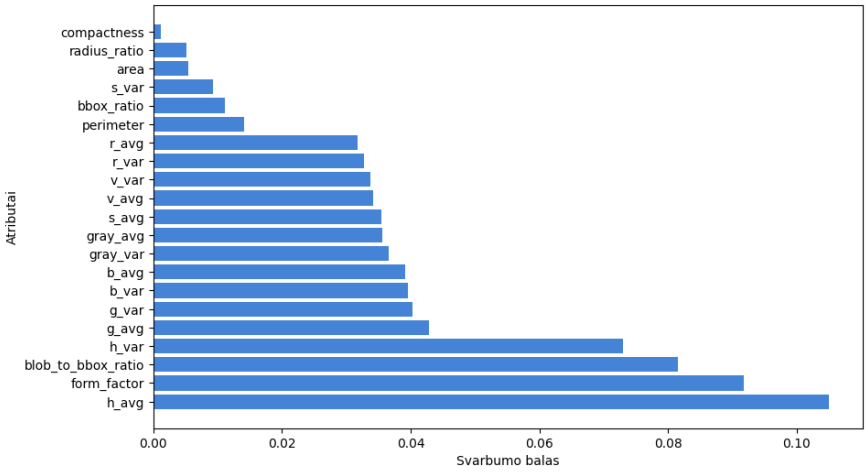
Parametrai	Reikšmės
n_neighbors	1, 2, 3, ... , 100
weights	uniform, distance
metric	manhattan, minkowski
algorithm	ball_tree, kd_tree, brute

Pirmajam modelio eksperimente atliekamas parametrų derinimas pasitelkus *scikit-learn* bibliotekos galimybes su visu atributų sąrašu. Pirmojo kodo implementacija vyko taikant kaimynų

skaičiaus (angl. *n_neighbors*) 1 - 100 diapazonu, svorio (angl. *weights*) ir atstumo matavimais (angl. *metric*) (2.2. skyrius). Optimaliausi nustatyti mokymo parametrai pirmame klasifikatoriaus duomenų mokyme naudojant 10 kryžminį patvirtinimą buvo Manheteno atstumo matavimas, kuris grįstas kaip grafiko taškų (koordinačių) vienos ašies skirtumas modulių sumai. Kiti geriausi mokymo parametrai - kaimynų skaičius 52 ir verčių svoris nepriklausomai nuo atstumo yra vienodas, tačiau gauti rezultatai nebuvo tobuli, mikroplastiko klasės tikslumas siekė 0.63, jautrumas 0.51 ir šių metrikų balanso vertė 0.56 (6 lentelė, Parametrų derinimas). Toliau siekiant geresnių rezultatų taikomi parametrų tinklelyje papildomi *scikit-learn* siūlomi parametrai: rutuliniai medžiai (angl. *ball_tree*), KD medžiai (angl. *KD_tree*), prievarta naudojami (angl. *brute force*) (palygina visus duomenų taškus tarpusavyje ieškant artimiausių kaimynų) ir mazgo dydis (angl. *leaf_size*), kuris turi tiesioginį poveikį prieš tai minėtiems medžiams. Gauti rezultatai nepakito, todėl gilesnė analizė apie šių parametrų veikimą nebuvo atliekami.

Antrame eksperimente ieškomas tinkamiausias atributų derinys su kuriais klasifikatorius gautų aukščiausius tinkamai aptiktų mikroplastiko klasės rezultatus. Pirmasis bandymas atliekamas naudojant geometrinius ir pilko fono atributus mokymui, taip bandant išvengti netolygių tarpusavyje anotuotų nuotraukų spalvų kontrastais. Kadangi mikroplastiko dalelė gali būti persidengus su dirbtinai sudarytu nuotraukos spalvos kontrastu. Gauti rezultatai nebuvo efektyvesni nei pirmajame eksperimente: tikslumas 0.66, jautrumas 0.49, F1 0.56. Kitas atributų pasirinkimas klasifikatoriaus mokymui taikomos tik spalvinės savybės, kurios tik pablogino mikroplastiko klasės rezultatus: tikslumas 0.58, jautrumas 0.42, F1 0.49. Tai reiškia, kad suderinus abejas kartu tiek spalvinės ir geometrinės savybės yra svarbios modelio mokymui.

Eksperimente taikomas *scikit-rebate* bibliotekos ReliefF metodas, naudojant parametrų tinklelį nustatytas efektyviausias kaimynų skaičius, šiam metodui. Žemiau pateiktas (17 pav.) atributų svarbumo grafikas kuris rodo, kaip skirtingi duomenų rinkinio atributai yra įvertinti pagal jų svarbą, remiantis ReliefF algoritmo rezultatais.



17 pav. Atributų svarbumo įvertinimas pasitelkus ReliefF metodą.

Palyginus su 15 pav. gauname nevienodus aukščiausius svarbumo atributus pvz. kaip objekto forma ir dominuojančios spalvos *h_avg* atributai. Klasifikatoriaus mokymui pasitelkus pirmuosius 4, 8 ar 15 geriausiai įvertintus atributus pagal 17 pav. gauti rezultatai nėra naudingesni nei pastarieji (tikslumas 0.61, jautrumas 0.44, F1 0.51). Kitame šio eksperimento bandyme taikyti gauti atsitiktinio miško vidutinio nevienodumo dydžio efektyviausi parametrai (15 pav.). Panaudoti pirmieji 9 atributai suteikė KNN modeliui efektyviausius rezultatus: tikslumas 0.65, jautrumas 0.54,

F1 0.59 (6 lentelė, Atributų svarbumo įvertinimas). Efektyviausi parametrai - kaimynų skaičius 62, Euklido atstumo matavimas ir svoris nepriklausomai nuo atstumo yra vienodas.

Trečiame K - artimiausio kaimyno klasifikatoriaus bandyme (6 lentelė, HNM) naudotas sunkių neigiamų pavyzdžių metodas, kuris 5.2. skyriuje aptarto atsitiktinio miško modelyje pateikė aukščiausius mikroplastiko objektų atpažinimo tikslumo rezultatus. Kiekvienos iteracijos metu (100 iteracijų) gautas KNN klasifikatoriaus metrikų vertės stabiliai svyruoja tarp dvejų pastovių modelio gaunamų verčių. Naudojant sunkių neigiamų pavyzdžių metodą pasiektas optimalus tikslumo naudingumo padidėjimas 0.72 tačiau dėl gauto mažesnio jautrumo 0.50 bendras modelio balansas nepakito.

Paskutiniame etape taikomas SMOTE metodas ir duomenų mažinimo metodas kaip ir 5.2. skyriuje atsitiktiniame miško klasifikatoriaus modelyje. Taikomi skirtingi klasių netikrų duomenų generavimų dydžiai nesuteikė modeliui pranašumo prieš kitus bandymus, o galutiniai rezultatai nepasiekė naudingesnio modelio našumo. Galutiniai šio eksperimento rezultatai (6 lentelė, Klasių balansavimas): tikslumas 0.44, jautrumas 0.90, F1 0.59. Efektyviausi parametrai - kaimynų skaičius 4, Manheteno atstumo matavimas, o svoris priklausomas nuo atstumo.

6 lentelė. KNN klasifikatoriaus įvertinimas atliekant skirtingus taikymus, pagal validavimo duomenų rinkinį.

Eksperimentas	Tikslumas	Jautrumas	F1
Parametrų derinimas	0.63	0.51	0.56
Atributų svarbumo įvertinimas	0.65	0.54	0.59
HNM	**0.72**	**0.50**	**0.59**
Klasių balansavimas	0.44	0.90	0.59

5.4. AdaBoost

Adaptyvaus stiprinimo klasifikatoriaus eksperimentų bandymų rezultatai pateikti 8 lentelėje. Žemiau pateiktoje 7 lentelėje yra nurodytos visos AdaBoost tyrime taikytos parametro tinklelio reikšmės.

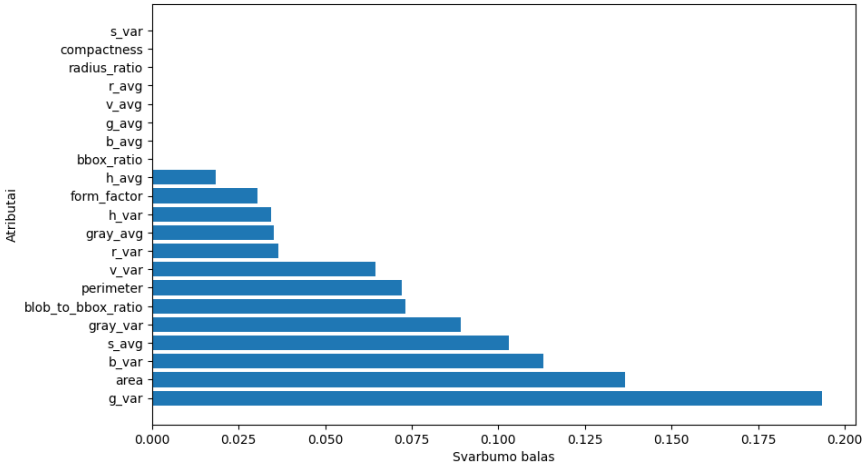
7 lentelė. AdaBoost klasifikatoriuje taikytos parametrų tinklelio reikšmės.

Parametrai	Reikšmės
estimator	1, 2, 3, 4, 5, 6, 7, 8, 9, 10
n_estimators	10, 50, 100, 200, 300, 400, 500
learning_rate	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1

Pirmuosiuose modelio mokymuose ieškomas geriausių parametrų derinys, jame naudojami klasifikatoriaus modelių kiekio skaičius (angl. *n_estimators*) ir mokymosi koeficientas (angl. *learning rate*). Modelių kiekio skaičius nurodo kiek atskirų vidinių klasifikatorių bus sukurta ir įtraukta į galutinį adaptyvaus stiprinimo modelį, o mokymosi koeficientas nustato kokia yra kiekvieno atskiro modelio įtaka galutiniam sprendimui. Mažesnis mokymosi koeficientas reiškia, kad kiekvieno modelio neteisingai klasifikuoti objektai turi mažesnį vertinimo svorį dėl to modelis vykdo atidesnį modelio mokymą. Didesnis koeficientas suteikia didesnį modelio svorį su kuriuo

skaičiavimai vyks greičiau, bet dėl to atsiranda modelio persimokymo riziką. Pirmieji eksperimento bandymų rezultatai su efektyviausiais rastais parametrais (mokymosi koeficientas 1, modelių kiekis 50): tikslumas 0.62, jautrumas 0.56, F1 0.58. Kitame modelio parametrų derinime taikytas *scikit-learn* bibliotekos siūlomas papildomas parametras - vertintojas (angl. *estimator*). Šį metodą AdaBoost klasifikatorius naudoja sprendimų medžio ir gilumo dydžio (angl. *max depth*) parametrui nurodyti, kurios numatytoji vertė yra 1. Tai reiškia, kad kiekvienas sprendimų kelmas (angl. *decision stumps*) gali atlikti ribotą klasifikavimo sprendimą remdamasis viena savybe, todėl įtraukiant papildomus kelmus visuose vidiniuose modeliuose įgaliname modelius detaliau analizuoti duomenis ir atpažinti tarpusavių ryšius. Antrajame bandyme gauti rezultatai efektyvesni (8 lentelė , Parametrų derinimas): tikslumas 0.63, jautrumas 0.65, F1 0.64. Optimizuotus rezultatus pavyko pasiekti su nustatytu šiuo geriausiu parametrų deriniu: medžių gylis 2, mokymosi svoris 0.4, o modelių kiekis 10.

Antrame eksperimente vykdoma svarbiausių atributų paieška pasitelkus nevienodumo pagrindo metodą (angl. *impurity based*), kuris sprendimų medžiuose (kelmuose) įvertina, kaip kiekvienas atributas padeda sumažinti klaidingai klasifikuotų atvejų skaičių. AdaBoost klasifikatoriuje įvertinama kiekvieno atributo svarba vidiniuose kelmuose ir remiantis jo indėliu per visą modelių iteraciją. Atributai, kurie nuosekliai prisideda prie klaidingai klasifikuotų atvejų mažinimo per visus medžius, gauna aukštesnį svarbos balą. Žemiau 18 pav. pateiktas šio pritaikyto metodo grafikas, kuriame atvaizduotas atributų svarbumas.



18 pav. Atributų svarbos atrinkimo metodo grafikas, teigiamos mikroplastiko klasės metrikų vertinimas.

Taikant savybių išskyrimo metodikas (15 pav., 17 pav., 18 pav.) atributų reikšmingumai skiriasi priklausomai nuo taikomo modelio ir metodo. Šį kartą pasitelkus šių grafikų duomenimis nepavyko išgauti naudingesnių rezultatų nei 8 lentelėje esančiais parametrų derinimo rezultatais. Buvo vykdomi skirtingų kiekių atributų mokymai, taip pat panaudotas geriausias atsitiktinio miško ir artimiausių kaimynų atitikusių atributu seka, tačiau rezultatai svyravo labai artimai arba jų naudingumas mažėjo. Geriausi šio antro eksperimento rezultatai pateikti 8 lentelėje - eksperimentas atributų svarbumo įvertinimas, kurių efektyviausi parametrai yra kai medžių gylis 2, mokymosi svoris 0.3, o modelių kiekis 10 .

Sekančiame šio klasifikatoriaus taikyme vykdomas HNM metodas (pasitelkus AdaBoost pirmo eksperimento efektyviausiu modeliu), kurį derinant su atsitiktiniu mišku ir KNN sukurti universalūs modeliai, gebantys identifikuoti ir atskirti sunkiai klasifikuojamus duomenų atvejus, o tai

padidino klasifikatorių tikslumą. Šiame bandyme klasifikatoriaus iteracijų modelio našumas vertės svyruoja dažniau nei KNN metode, kas 6 iteracijas gaunamas modelio geriausias balansas su aukščiausia tikslumo verte, kai tikslumas 0.70, jautrumas 0.63, F1 66 (lentelė 8, HNM).

Paskutiniame bandyme vėl taikomas duomenų mažinimas ir sintetinis jų didinimas taikant SMOTE, tačiau rezultatai nepaisant klasifikatoriaus tipo nesikeičia (lentelė 8, Klasių balansavimas), mažas tikslumas 0.40, o jautrumas aukštas 0.90. Šiuos rezultatus pavyko pasiekti su nustatytu šiuo geriausiu parametru deriniu: medžių gylis 2, mokymosi svoris 0.4, o modelių kiekis 100. Taip vyksta dėl to nes duomenų mažinimas neigiamoje klasėje pašalina svarbius išskirtinius duomenis kurie modelio mokymui yra vertingi, o dirbtinai sukurti duomenys nėra tikrieji duomenys dėl kurių modelis negali įvertinti tikrąjį teigiamų atvejų pasiskirstymą.

8 lentelė. Adaptyvaus stiprinimo klasifikatoriaus įvertinimas atliekant skirtingus taikymus, pagal validavimo duomenų rinkinį.

Eksperimentas	Tikslumas	Jautrumas	F1
Parametrų derinimas	0.63	0.65	0.64
Atributų svarbumo įvertinimas	0.63	0.64	0.64
HNM	**0.70**	**0.63**	**0.66**
Klasių balansavimas	0.40	0.90	0.56

5.5. XGBoost

Ekstremalaus gradiento stiprinimo klasifikatoriaus eksperimentų rezultatai pateikti 10 lentelėje. Žemiau pateiktoje lentelėje 9 yra nurodytos visos XGBoost tyrime taikytos parametro tinklelio reikšmės.

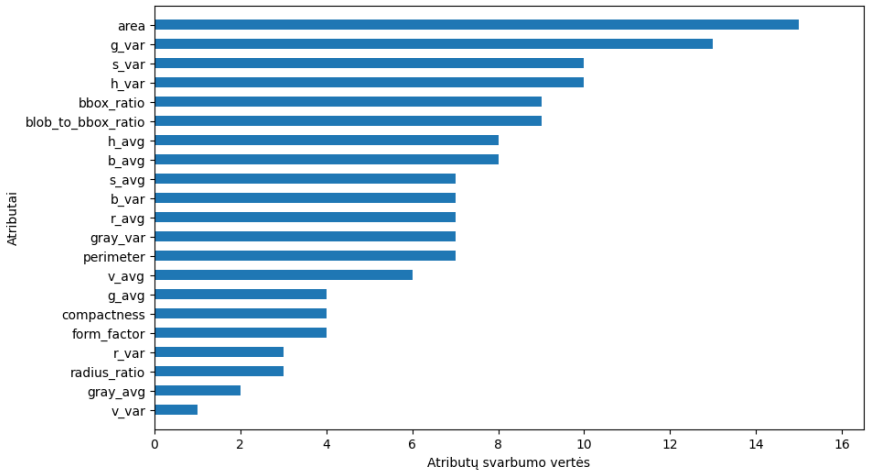
9 lentelė. AdaBoost klasifikatoriuje taikytos parametrų tinklelio reikšmės.

Parametrai	Reikšmės
n_estimators	10, 25, 30, 50, 100, 150
learning_rate	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1
max_depth	1, 2, 3, 4, 5, 6, 10
gamma	0, 0.1, 0.2 , 0.3, 0.5, 0.7, 1
subsample	0.6, 0.7, 0.8, 1.0
colsample_bytree	0.5, 0.6, 0.7, 0.8, 1.0
reg_lambda	0, 0.5, 1, 1.5, 2
reg_alpha	0, 0.1, 0.5, 1, 2
booster	gblinear, dart, tree
normalize_type	tree, forest
rate_drop	0, 0.1, 0.5, 1
one_drop	0, 1
skip_drop	0, 0.1, 0.5, 1

Pirmuosiuose šio modelio mokymuose ieškomi efektyviausi parametrai turimam duomenų rinkiniui taikant medžių stiprinimo metodą (angl. *tree booster*). Medžių stiprinimas yra tradicinis

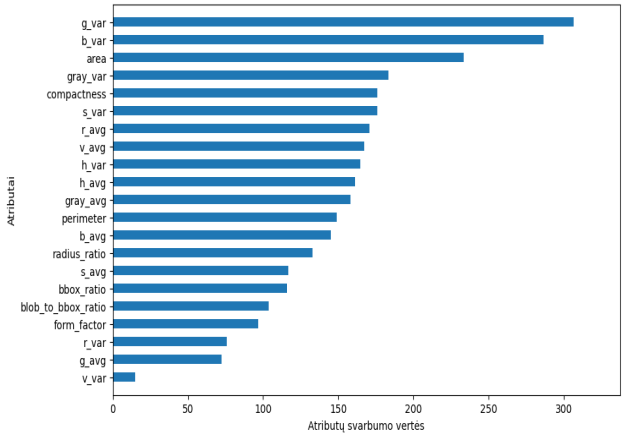
XGBoost metodo pagrindas, jo priskirti parametrai sukuria sudėtingus medžių modelius. Pirmajam bandyme naudoti tokie patys parametrai kaip ir adaptyviajame stiprinimo klasifikatoriuje: modelių kiekio skaičius, mokymo koeficientas ir medžių gylis. Šio tyrimo klasifikatoriaus maksimalus medžių gylis 2 suteikė efektyviausius rezultatus (tikslumas: 0.65, jautrumas 0.70, F1 0.68), kartu su mokymo koeficientu 0.5 ir modelių kiekiu 10 (rezultatai pateikti lentelėje 10, Parametrų derinimas). Sekančiame bandyme įtraukti medžių stiprinimo papildomi XGBoost bibliotekos parametrai, tai γ minimalus medžių mazgų skaidymo skaičius, L2 ir L1 reguliarizacijos parametrai. Optimaliausios parametrų vertės, atrinktos naudojant parametrų tinklelį, rodo, kad kai kurie parametrai, kaip λ= 1, α= 0, yra pasirinktos kaip numatytosios vertės. Kiti parametrai γ= 0.5, mokymosi koeficientas 0.5, maksimalus medžių gylis 2, svorių suma 2, ir modelių kiekis 30 pateikė minimaliai prastesnius rezultatus: tikslumas 0.63, jautrumas 0.69, F1 0.66. Likusieji stiprinimo metodai tokie kaip atmetimo stiprinimo (angl. *dart booster*) ir linijinio stiprinimo (angl. *linear booster*) taikyti eksperimente būdai suteikė skirtingas naudingo našumo rezultatus. Pagrindinė atmetimo stiprinimo savybė yra atsitiktinis medžių atmetimas mokymosi proceso metu. Atmetimo dažnumą nurodo trys parametrai: dažnumo atmetimas (angl. *rate_drop*) nustato kokia tikimybe medžiai bus atmetami, išmetimo kiekis (angl. *one_drop*) nustato ar bent vienas medis turi būti atmetamas ir atmetimo praleidimo (angl. *skip_drop*) tikimybė, visi šie parametrai vykdomi po kiekvienos medžių sukūrimo iteracijos. Baigus visas iteracijas, galutinėje modelio prognozės vertinimui įtraukiami visi praleisti modelių medžiai [7]. Šis taikomas sprendimas nesuteikė efektyvesnių rezultatų nei medžių stiprinimo metodas: tikslumas 0.64, jautrumas 0.69, F1 0.66. Linijinis stiprinimo metodas pasiekė prastus rezultatus, F1 vertė siekė tik 0.37, o tikslumas 0.53.

Antrame šio klasifikatoriaus eksperimente vykdomas geriausių atributų paieška taikant skirtingus XGBoost atributų svarbumo skaičiavimo metodus, tai svorių, dengimo (angl. *cover*) ir gavimo (angl. *gain*) įvertinimai. Svorių metodas (19 pav.) skaičiuoja, kiek kartų kiekvienas atributas naudojamas visuose modelio sukuriamuose medžiuose. Tačiau šis metodas nesuteikia informacijos apie modelio atliktą prognozę. Dengimo metodas (20 pav. (a)) įvertina, kiek vidutiniškai duomenų taškų kiekvienas atributas reprezentuoja medžių skaidymo mazguose, kiek kartų atributas atsiranda modelio medyje tiek jo indėlis įtraukiamas į modelio prognozę. Gavimo metodas (20 pav. (b)) įvertina atributo indėlį į modelio sprendimų tikslumą, kur atributas sudarė prognozavimo vertinimą modelio tikslumui visuose medžių skaidymo procesuose [7]. Atributų pagal jų svarbumo svorių skaičiavimus pasitelkus efektyviausius 14 atributų (vertės nuo 8 balų) iš 19 pav. pateikia nedidelį modelio naudingumo nukritimą: tikslumas 0.63, jautrumas 0.71, F1 0.66. Bandant mažinti atributų skaičių atitinkamai nuo mažiausiai svarbaus, F1 vertė viršija pirmojo eksperimento rezultatą kai lieka tik 4 aukščiausi įvertinti atributai (F1 0.69), o tikslumas išlieka svyruodamas tarp 0.63 ir 0.62 verčių. Jautrumas tuo tarpu išauga iki 0.76, tai reiškia svorių didinimo metodas šiam duomenų rinkiniui nėra naudingas, kadangi siekiama turėti kuo didesnė teigiamos klasės tikslumą.

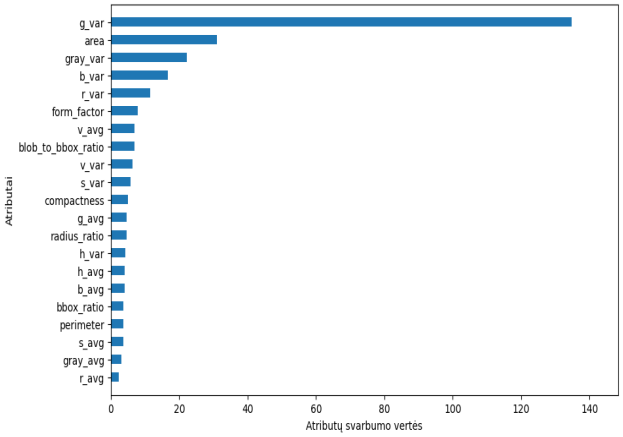


19 pav. XGBoost atributų svarbos grafikas pagal svorius.

Kiti ekstremalaus gradiento stiprinimo atributų svarbumo skaičiavimo matavimai pateikė kitokius rezultatus (20 pav.). Dengimo metodas su 12 atributų (vidurkių vertės nuo 150) pateikė efektyviausius tikslumo metrikų įverčius (tikslumas 0.66, jautrumas 0.74, F1 0.70) kai mokymosi koeficientas 0.1, medžių gylis 4, o modelių kiekis 50. Bandant mažinti atributų skaičių, rezultatai prastėjo. Gavimo metodas rodantis kiekvieno atributo indėlį į modelio tikslumą, įvertino, kad ypač svarbus yra g_var atributas. Atliekant bandymus su skirtingais atributų kiekiais, galutiniai rezultatai nėra naudingesni nei dengimo metodo, jie siekia panašius metrikų įverčius kaip svorių metodo rezultatų. Taip yra dėl nesubalansuotų klasių, kurių bendras modelių tikslumas yra klaidinantis šiuo atveju mažesnės teigiamos klasės atveju. Optimaliausi dengimo rezultatai pateikti 10 lentelėje, Atributų svarbumo įvertinimas.



(a) Atributų svarbumas pagal dengimą.



(b) Atributų svarbumas pagal gavimą.

20 pav. Ekstremalus gradiento stiprinimo atributų svarbos grafikai.

Trečiasis eksperimentas su ekstremalaus gradiento stiprinimu, kuriame taikomas sunkių neigiamų pavyzdžių metodas pagal antro eksperimento efektyviausią modelį, suteikė naudingus galutinius modelio rezultatus. Modelio mokymo iteracijos rezultatų spektras daug didesnis nei praeitų eksperimentų. Todėl prireikė didesnio XGBoost HNM iteracijų skaičiaus 200. Nepaisant didesnio iteracijų skaičiaus metrikų rezultatai išlieka stabilūs, o tai rodo, kad modelis nepersimoko. Gauti efektyviausi rezultatai (10 lentelėje, HNM): tikslumas 0.73, jautrumas 0.70 ir F1 0.71. Lyginant šiuos rezultatus su antrame eksperimente užfiksuotomis naudingiausiomis vertėmis, pastebimas

aiškus jautrumo sumažėjimas bei tikslumo išaugimas. Tai rodo, kad sunkių neigiamų pavyzdžių metodas modeliui leidžia efektyviai vertinti neigiamus atvejus, mažindamas klaidingai teigiamų atpažinimų skaičių.

10 lentelė. XGBoost klasifikatoriaus įvertinimas atliekant skirtingus taikymus, pagal validavimo duomenų rinkinį.

Eksperimentas	Tikslumas	Jautrumas	F1
Parametrų derinimas	0.65	0.70	0.68
Atributų svarbumo įvertinimas	0.66	0.74	0.70
HNM	**0.73**	**0.70**	**0.71**

5.6. Galutinis modelių vertinimas

Analizuojant visų klasifikatorių atliktus eksperimentus, pradiniai bandymai visuose metodų taikymuose vyko ieškant tinkamiausių modelių parametrų, kurie suteiktų aukščiausius tikslumo ir jautrumo rezultatus. Tai svarbu, atsižvelgiant į tai, kad ne mikroplastiko klasė sudaro daugiau nei 10 kartų didesnę dalį duomenų rinkinio ir dėl to modelių daugiau atstovų turinčios klasės rezultatai artimi 1. Dėl to pagrindinis dėmesys skirtas teigiamos mikroplastiko klasės naudingumo gerinimui, neatsižvelgiant į bendrą modelio tikslumą. Kiekvieno modelio parametrų derinimas taikant skirtingus duomenų rinkinius pasitelkus stratifikuotą 10 kryžminį patvirtinimą leido modeliams adaptuotis prie įvairių duomenų bruožų. Atsitiktinio miško klasifikatoriaus atveju, parametrų derinimas ir klasifikatoriaus integruotas atributų svarbumo vertinimas naudingas. Atrinkus tinkamiausius parametrus, medžių gylį, skaidymo kriterijus, duomenų limito vertes ir įgyvendinus MDI metodą buvo sėkmingai pasiektas modelio efektyvumo tobulėjimas ir sumažintas klaidingai aptiktų objektų teigiamų atvejų kiekis. Paprastesniems modeliams tokiems kaip KNN ir AdaBoost taikant parametrų derinimą pavyko pagerinti klasifikavimo naudingumo vertes. Tačiau jų galimybės efektyviai prisitaikyti prie duomenų yra ribotesnės nei atsitiktinio miško. Tinkamas parametrų derinimas nors ir gali šiek tiek padidinti modelio tikslumą, tačiau trūkumas yra tas, kad šie modeliai nesuteikia galimybės įvertinti atributų svarbumus modelių mokymuose. Dėl šios priežasties vykdomi išoriniai atributų svarbumo vertinimo metodai: ReliefF ir *Impurity based*. Tačiau šie modeliai nepasiekė panašaus efektyvumo lygio, kaip atsitiktinio miško ar XGBoost modeliai, kurie integruoja atributų svarbos analizę. Taikant KNN modeliui atsitiktinio miško MDI metodo rastus efektyviausius atributus padėjo modeliui gauti naudingesnius rezultatus nei jam naudojant ReliefF metodą. Toks sprendimas Adaboost ir XGBoost modeliuose nepasiteisino. Pats XGBoost metodas išsiskiria savo parametrų derinimų gausa, kas suteikia modeliui galimybę prisitaikyti prie skirtingų, įvairių duomenų rinkinių. Norint pasiekti optimaliausius rezultatus užteko derinti šiuos parametrus: medžių gylis, mokymosi koeficientas ir modelių kiekis. Taip pat XGBoost integruotos atributų svarbumo vertinimo technikos suteikė detalų vertinimą apie kiekvieno atributo įtaką modelio sprendimų procese. Dengimo metodas turėjęs didžiausią indelį į modelio teigiamos klasės tikslumo gerinimą. Kitas svarbus klasifikavimo tyrimo etapas sunkių neigiamų pavyzdžių metodo taikymas modeliams, kurie optimizuoti atsižvelgiant į parametrų derinimą ir savybių svarbos išskyrimą. Šis metodas padidino modelių gebėjimus mažinti klaidingai aptiktų objektų teigiamų atvejų skaičių. HNM taikymas visuose modeliuose parodė pastebimą tikslumo padidėjimą, kartu sumažėjusį jautrumą. Tai rodo, kad modeliai tapo efektyvesni atpažįstant tikruosius neigiamus mikroplastiko atvejus. Pasitelkus duomenų balansavimo metodus (SMOTE ir duomenų suprastinimą) pavyko suvienodinti klasių duomenų rinkinio duomenis. Tačiau toks modelio mokymas neleido pasiekti efektyvių rezultatų. Papildomai taikant HNM metodą ant dirbtinai sukurtų ar pašalintų duomenų pastebėta, kad ši kombinacija sukelia per didelį modelio prisitaikymą prie mokymo duomenų.

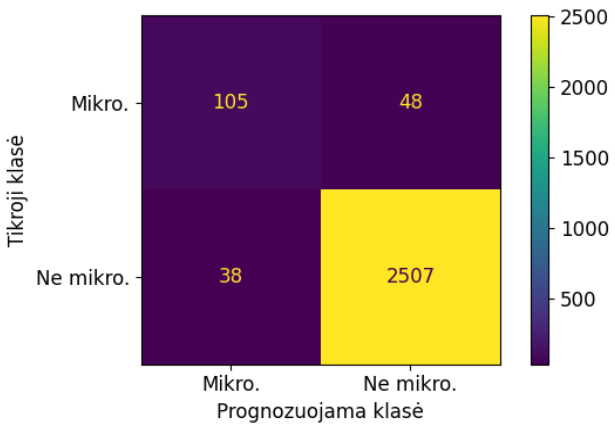
Žemiau 11 lentelėje pateikti galutiniai mikroplastiko klasifikavimo rezultatai, parodantys kaip klasifikavimo modeliai prognozuoja mikroplastiką dar nematytuose duomenyse, testavimo duomenų rinkinyje. Šie galutiniai rezultatai išgaunami iš kiekvieno taikyto klasifikatoriaus efektyviausių rezultatų, kuriuose naudojamas sunkių neigiamų pavyzdžių metodas. Pirmasis taikytas atsitiktinis miško klasifikatorius pasiekė aukščiausius rezultatus tarp visų taikytų metodų (tikslumas 0.73, jautrumas 0.69, F1 0.71), pasižymėdamas efektyviu balansu tarp tikslumo ir jautrumo. Sekantis KNN klasifikatorius pasiekė aukštą tikslumą 0.72, o jautrumas dėl modelio paprastumo ar nepakankamo duomenų savybių atskleidimo išliko žemas 0.52, galutinis F1 įvertis 0.60. AdaBoost

11 lentelė. Mikroplastiko klasifikavimo galutiniai rezultatai pagal testavimo duomenų rinkinį.

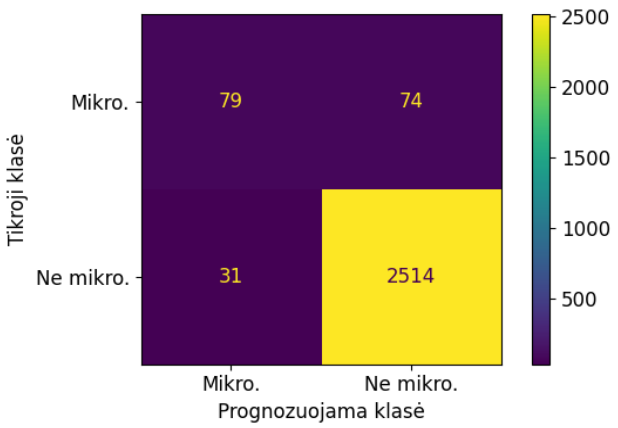
	Mikroplastiko klasė		
Klasifikatorius	Tikslumas	Jautrumas	F1
Atsitiktinis miškas	**0.73**	**0.69**	**0.71**
KNN	0.72	0.52	0.60
AdaBoost	0.68	0.65	0.67
XGBoost	0.71	0.65	0.68

modelis pasiekė tikslumą 0.68, jautrumą 0.65 ir F1 įvertį 0.67, rodydamas subalansuotą veikimą, bet ne tokį efektyvų kaip atsitiktinis miškas. XGBoost modelio galutiniai rezultatai testavimo rinkinyje yra prastesni (tikslumas 0.71, jautrumas 0.65, F1 0.68) nei gauti iš validavimo duomenų rinkinio.

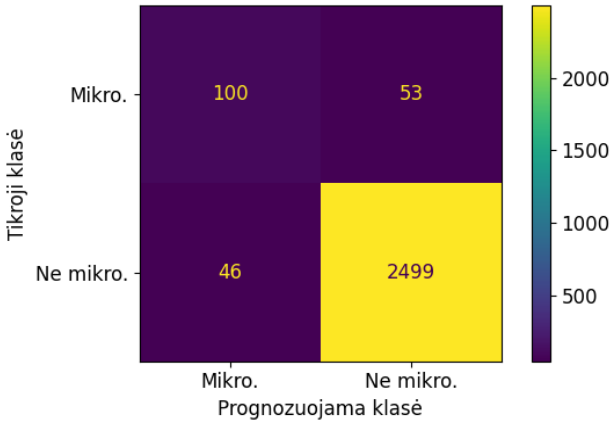
Kitas svarbus klasifikatorių modelių veikimo palyginimas atliekamas pasinaudojant painiavos matrica (angl. *confusion matrix*), kuri parodo kaip tiksliai modelis atpažįstą kiekvieną klasę. Kiekviena matrica žemiau pavaizduota 21 pav. turi 4 pagrindinius skaičius nurodančius: teisingai prognozuojamą mikroplastiką (TP), klaidingai prognozuojamą mikroplastiką (FP), teisingai prognozuojamą ne mikroplastiką (TN) ir klaidingai ne prognozuojamą mikroplastiką (FN). Atsitiktiniame miške (21 pav. (a)) modelis efektyviausiai prognozuoja teigiamas mikroplastiko klases. Taip pat atsitiktinis miškas sudaro mažiausią klaidingai teigiamų atvejų (FP) skaičių iš visų klasifikatorių, kuris ženkliai mažesnis lyginant su KNN modelyje (21 pav. (b)) esančiais 74 klaidingai prognozuojamais mikroplastiko atvejais. Tačiau KNN modelis pateikia geresnį klaidingai neigiamos klasės atvejų (FN) skaičių 31, o atsitiktinio miško modelis su 38 atvejais, tačiau pernelyg didelis FP sumažina KNN modelio balansą tarp prognozuojamų mikroplastiko ir ne mikroplastiko klasių atvejų skaičių. Dėl to atsitiktinis miškas išlieka efektyvesnis bendrame modelių vertinime, nes svarbu ne tik teisingai atpažinti mikroplastikai bet ir mažesnis FN taip pat išvengiant FP atvejų, užtikrinant efektyvesnius rezultatus. Pagal AdaBoost painiavos matricos (21 pav. (c)) rezultatus galime pamatyti, kad modelis sugeba išlaikyti balansą tarp FP 53 ir FN 46, siekiant TP 100 prognozuojamų atvejų, o tai rodo, kad modelis yra linkęs praleisti mažiau tikrųjų mikroplastiko atvejų nei KNN. XGBoost modelis (21 pav. (d)) panašiai jautrus klaidingai teigiamiems atvejams kaip AdaBoost, bet su šiek tiek didesniu FP skaičiumi. Vis dėlto, XGBoost ir AdaBoost nors ir turi panašius modelio prognozavimo atvejų skaičius atpažįstant efektyviai tikruosius mikroplastiko atvejus, tačiau nepranokstą atsitiktinio miško klasifikatoriaus rezultatų. Akivaizdžiai visuose modeliuose pastebimas aukštas TN atvejų skaičius, kur dėl duomenų nevienodumo šie skaičiai yra tendencingi, nes modeliai yra linkę teisingai atpažinti būtent dominuojančią klasę.



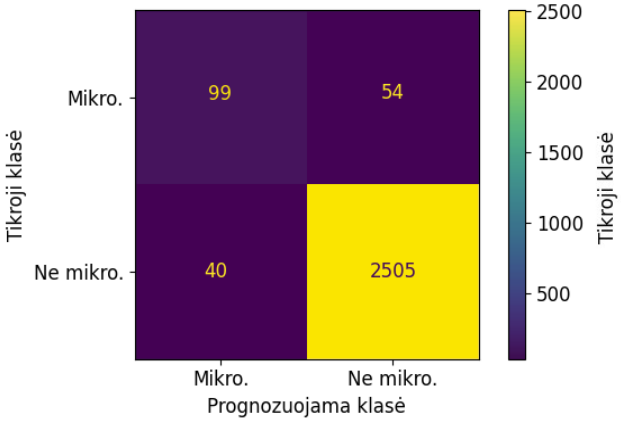
(a) Atsitiktinio miško klasifikatoriaus.



(b) KNN klasifikatoriaus.



(c) Adaboost klasifikatoriaus.



(d) XGBoost klasifikatoriaus.

21 pav. Mikroplastiko klasės testavimo rinkinio duomenų painiavos matricos.

Išvados ir rekomendacijos

Šiame darbe taikomi binarizacijos metodai nevienodai apšviestose mikroskopijos nuotraukose kartu su mašininio mokymosi algoritmais skirti mikroplastiko dalelių efektyviam aptikimui.

- Tyrimo metu atlikti eksperimentai naudojant Otsu ir Sauvola binarizacijos metodus parodė, kad Sauvola metodas yra efektyvesnis, nes aptinka didesnį tikrųjų mikroplastiko dalelių skaičių, tačiau kartu išauga ir klaidingai teigiamų mikroplastiko aptikimo atvejų, dėl fiksuojamų mikroskopo filtro dėmių. Be to, abu metodai, netiksliai įvertinę mikroplastiko koncentraciją, vienalytės dalelės yra skaidomos į kelias mažesnes dalis, kas taip pat prisideda prie klaidingai teigiamų mikroplastiko aptikimo atvejų.
- Pasirinktas optimalesnis Sauvolos metodas, kuris buvo taikomas su morfologiniais sprendimais, efektyviai sumažino klaidingai teigiamų mikroplastiko dalelių aptikimo atvejų skaičių, tačiau išliko ganėtinai mažas aptiktų tikrųjų mikroplastiko objektų tikslumas (angl. *precision*).
- Mašininio mokymosi metodų taikymas, atsižvelgiant į duomenų klasės disbalansą, leido pasiekti efektyvesnius rezultatus, nustatant optimalius parametrus ir efektyviausius specifinius mikroplastiko atributus. Ypatingai svarbus sunkių neigiamų pavyzdžių (angl. *Hard negative mining*) metodas, kuris sumažino klaidingai teigiamų objektų atpažinimo skaičių naudojamiems atsitiktinio miško, KNN, AdaBoost ir XGBoost klasifikatoriams. Efektyviausius rezultatus nevienodai apšviestose mikroskopijos nuotraukose pateikė atsitiktinio miško klasifikatorius.

Mašininio mokymosi nuoseklus modelių tobulinimas ir jų taikymas specifiniams vaizdams, leido pasiekti optimalius rezultatus. Tačiau lyginant su esamais tokio tipo projektais, galutiniai rezultatai prastesni. Taip yra dėl to, kad esamuose tyrimuose mikroplastiko nuotraukos dažniausiai neturėjo šviesos ir kontrasto skirtumų, ar mikroskopo filtro dėmių. Todėl vykdant tokio pobūdžio projektą svarbu išgauti, kuo skaidresnius ir detalesnius vaizdus leidžiančius binarizacijos metodams efektyviau aptikti tik mikroplastiko daleles, mažinant triukšmo ir netikslumo poveikį sprendimų priėmimo procesuose.

Ateities tyrimų planas

Ateities tyrimo plane būtų tikslinga ištirti ir spręsti tarpusavyje persidengiančių mikroplastiko dalelių problemą, kadangi po vaizdo binarizacijos mikroplastiko objektų kontūrai susijungia ir gaunama vienalytė dalis, nors iš tikrųjų tai gali būti skirtingos tipo dalelės. Šiame projekte naudotame duomenų rinkinyje pasitaikė tik viena tokia situacija, kuri neturėjo didelės įtakos bendram galutiniam rezultatui. Vis dėlto, jei duomenų rinkinį sudarytų didesnis kiekis vaizdų, kur mikroplastiko objektai persidengia tarpusavyje, tuomet galutiniai rezultatai prastėtų. Todėl svarbu nustatyti ir taikyti pažangius vaizdų analizės metodus, kurie leistų efektyviau atskirti persidengiančias daleles. Tokios situacijos sprendimui gali būti pritaikytas vandens linijų (angl. *Watershed*) segmentavimo metodas, kuris leidžia efektyviai atskirti persidengiančias daleles [13]. Šis metodas konvertuoja vaizdą į topografinę formą, kurioje objektų briaunos išskiriamos remiantis šviesumo, kontrasto ir spalvų skirtumais. Taikant šią dalelių segmentaciją, padidintume tikrųjų teigiamų (TP) objektų atvejų skaičių, užtikrindami aukštesnį tikslumo ir jautrumo lygį mikroplastiko dalelių aptikimo užduotyje.

Sekančiame tyrimo etape svarbu toliau vykdyti mikroplastiko dalelių aptikimų analizę, taikant sudėtingesnius neuroninius tinklus dalelių klasifikavime ir U-Net dalelių segmentavime metodus. Šie metodai pasižymėjo efektyvesniais rezultatais nei Sauvola ir mašininio mokymosi modeliai pagal ištirtus darbus [18] [33].

Literatūros šaltiniai

[1] Daniel Berrar et al. Cross-validation., 2019.

[2] Ekaba Bisong and Ekaba Bisong. Google colaboratory. *Building machine learning and deep learning models on google cloud platform: a comprehensive guide for beginners*, pages 59--64, 2019.

[3] Leo Breiman. Bagging predictors. *Machine learning*, 24:123--140, 1996.

[4] Leo Breiman. Random forests. *Machine learning*, 45:5--32, 2001.

[5] John Canny. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, (6):679--698, 1986.

[6] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321--357, 2002.

[7] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785--794, New York, NY, USA, 2016. ACM.

[8] Bowen Cheng, Ross Girshick, Piotr Dollár, Alexander C Berg, and Alexander Kirillov. Boundary iou: Improving object-centric image segmentation evaluation. pages 15334--15342, 2021.

[9] Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21--27, 1967.

[10] Tamara S Galloway and Ceri N Lewis. Marine microplastics spell big problems for future generations. *Proceedings of the national academy of sciences*, 113(9):2331--2333, 2016.

[11] Estevão S Gedraite and Murielle Hadad. Investigation on the effect of a gaussian blur in image filtering and segmentation. In *Proceedings ELMAR-2011*, pages 393--396. IEEE, 2011.

[12] Roland Geyer. A brief history of plastics. *Mare plasticum-the plastic Sea: Combatting plastic pollution through science and art*, pages 31--47, 2020.

[13] Rafael C Gonzalez. *Digital image processing*. Pearson education india, 2009.

[14] Aaron Hertzmann, David Fleet, and Marcus Brubaker. Machine learning and data mining lecture notes. *University of Toronto Version*, 134, 2012.

[15] Michael I Jordan and Tom M Mitchell. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255--260, 2015.

[16] Xiao Li, Yu Wang, Sumanta Basu, Karl Kumbier, and Bin Yu. A debiased mdi feature importance measure for random forests. *Advances in Neural Information Processing Systems*, 32, 2019.

[17] Javier Lorenzo-Navarro, Modesto Castrillón-Santana, May Gómez, Alicia Herrera, and Pedro A Marín-Reyes. Automatic counting and classification of microplastic particles. In *ICPRAM 2018-Proceedings of the 7th International Conference on Pattern Recognition Applications and Methods*, 2018.

[18] Javier Lorenzo-Navarro, Modesto Castrillón-Santana, Elena Sánchez-Nielsen, Borja Zarco, Alicia Herrera, Ico Martínez, and May Gómez. Deep learning approach for automatic microplastics counting and classification. *Science of the Total Environment*, 765:142728, 2021.

[19] Javier Lorenzo-Navarro, Modesto Castrillon-Santana, Enrico Santesarti, Maria De Marsico, Ico Martínez, Eugenio Raymond, May Gomez, and Alicia Herrera. Smacc: a system for microplastics automatic counting and classification. *IEEE Access*, 8:25249--25261, 2020.

[20] Anqi Mao, Mehryar Mohri, and Yutao Zhong. Cross-entropy loss functions: Theoretical analysis and applications. *arXiv preprint arXiv:2304.07288*, 2023.

[21] Wan Azani Mustafa, Haniza Yazid, and Mastura Jaafar. An improved sauvola approach on document images binarization. *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)*, 10(2):43--50, 2018.

[22] Kemal Oksuz, Baris Can Cam, Sinan Kalkan, and Emre Akbas. Imbalance problems in object detection: A review. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3388--3415, 2020.

[23] OpenCV. Open source computer vision library, 2015.

[24] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, 9(1):62--66, 1979.

[25] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825--2830, 2011.

[26] Tamar Peli and David Malah. A study of edge detection algorithms. *Computer graphics and image processing*, 20(1):1--21, 1982.

[27] Felipe Pezoa, Juan L Reutter, Fernando Suarez, Martín Ugarte, and Domagoj Vrgoč. Foundations of json schema. pages 263--273, 2016.

[28] Philipp Probst, Marvin N Wright, and Anne-Laure Boulesteix. Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: data mining and knowledge discovery*, 9(3):e1301, 2019.

[29] P Ravi and A Ashokkumar. Analysis of various image processing techniques. *International Journal of Advanced Networking and Applications*, 8(5):86--89, 2017.

[30] Peter G Ryan. A brief history of marine litter research. *Marine anthropogenic litter*, pages 1--25, 2015.

[31] Jaakko Sauvola and Matti Pietikäinen. Adaptive document image binarization. *Pattern recognition*, 33(2):225--236, 2000.

[32] Abdenour Sehad, Youcef Chibani, Mohamed Cheriet, and Yacine Yaddaden. Ancient degraded document image binarization based on texture features. In *2013 8th International Symposium on Image and Signal Processing and Analysis (ISPA)*, pages 189--193. IEEE, 2013.

[33] Bin Shi, Medhavi Patel, Dian Yu, Jihui Yan, Zhengyu Li, David Petriw, Thomas Pruyn, Kelsey Smyth, Elodie Passeport, RJ Dwayne Miller, et al. Automatic quantification and classification of microplastics in scanning electron micrographs via deep learning. *Science of The Total Environment*, 825:153903, 2022.

[34] Piotr Skalski. Make Sense. 2019.

[35] Jingwen Sun, Weixing Du, and Niancai Shi. A survey of knn algorithm. *Information Engineering and Applied Computing*, 1, 05 2018.

[36] Satoshi Suzuki et al. Topological structural analysis of digitized binary images by border following. *Computer vision, graphics, and image processing*, 30(1):32--46, 1985.

[37] Ryan J Urbanowicz, Melissa Meeker, William La Cava, Randal S Olson, and Jason H Moore. Relief-based feature selection: Introduction and review. *Journal of biomedical informatics*, 85:189--203, 2018.

[38] Ryan J. Urbanowicz, Randal S. Olson, Peter Schmitt, Melissa Meeker, and Jason H. Moore. Benchmarking relief-based feature selection methods. arXiv e-print. https://arxiv.org/abs/1711.08477, 2017.

[39] Guido Van Rossum and Fred L Drake Jr. *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam, 1995.

[40] Daniela Witten and Gareth James. *An introduction to statistical learning with applications in R*. springer publication, 2013.

[41] Hong Xuan, Abby Stylianou, Xiaotong Liu, and Robert Pless. Hard negative examples are hard, but useful. In *Computer Vision--ECCV 2020: 16th European Conference, Glasgow, UK, August 23--28, 2020, Proceedings, Part XIV 16*, pages 126--142. Springer, 2020.