

ŽILINSKÁ UNIVERZITA V ŽILINE

Fakulta riadenia a informatiky

AUTOREFERÁT

DIZERTAČNEJ PRÁCE

Žilina, 2024

Mgr. Samuel Molčan

Žilinská univerzita v Žiline
Fakulta riadenia a informatiky

Mgr. Samuel Molčan

Autoreferát dizertačnej práce

Diagnostika krvi pomocou strojového učenia a simulácií

na získanie akademického titulu **„philosophiae doctor“** (v skratke **PhD.**)
v študijnom programe doktorandského štúdia
aplikovaná informatika

v študijnom odbore:
informatika

Žilina, apríl 2024

Dizertačná práca bola vypracovaná v dennej forme doktorandského štúdia na Katedre softvérových technológií na Fakulte riadenia a informatiky Žilinskej univerzity v Žiline.

Predkladateľ: Mgr. Samuel Molčan
Katedra softvérových technológií
Fakulta riadenia a informatiky
Žilinská univerzita v Žiline

Školiteľ: doc. RNDr. Katarína Bachratá, PhD.
Katedra softvérových technológií
Fakulta riadenia a informatiky
Žilinská univerzita v Žiline

Oponent: prof. RNDr. Vladimír Janiš, CSc.
Katedra matematiky UMB Tajovského 40, 97401 Banská Bystrica

Oponent: doc. Mgr. Ondrej Šuch, PhD.
Inštitút matematiky a informatiky MÚ SAV v Banskej Bystrici.
Dumbierska 1. 974 11 Banská Bystrica

Autoreferát bol rozoslaný dňa: 25. 6. 2024

Obhajoba dizertačnej práce sa koná dňa 21.8.2024 o 9:00 hod. pred komisiou pre obhajobu dizertačnej práce schválenou pracovnou skupinou odborovej komisie v študijnom odbore **informatika** v študijnom programe **aplikovaná informatika**, vymenovanou dekanom Fakulty riadenia a informatiky Žilinskej univerzity v Žiline.

prof. Ing. Ľudmila Jánošíková, PhD.
predsedníčka pracovnej skupiny odborovej komisie
v študijnom odbore **informatika**
v študijnom programe **aplikovaná informatika**

Fakulta riadenia a informatiky
Žilinská univerzita
Univerzitná 8215/1
010 26 Žilina

• Abstrakt

Práca skúma inovatívne metódy diagnostiky porúch súvisiacich s krvou pomocou kombinácie techník strojového učenia a simulácií, najmä so zameraním na analýzu červených krviniek prostredníctvom simulačného modulu ESPResSo. Tento prístup umožňuje hlboký ponor do pochopenia správania červených krviniek za rôznych fyziologických podmienok generovaním numerických výstupov zo simulácií. Cieľom je využiť modely strojového učenia, najmä neurónové siete a náhodné lesy, na analýzu týchto výstupov na klasifikáciu červených krviniek na základe zmien v ich elasticite, čo môže naznačovať prítomnosť porúch krvi. Táto nová metodológia sa snaží preklenúť priepasť medzi tradičnými diagnostickými technikami a obrovským potenciálom výpočtovej analýzy a ponúka presnejší a efektívnejší diagnostický nástroj.

Hlavným cieľom práce je presahovať rámec numerickej analýzy údajov transformáciou výstupov simulácie do dynamických vizuálnych reprezentácií. Zahŕňa to generovanie videí na zachytenie správania červených krviniek v toku krvi, ktoré sa ďalej spracúvajú modelmi strojového učenia na účely klasifikácie. Tento duálny prístup, ktorý zahŕňa numerické aj vizuálne údaje, má za cieľ zlepšiť diagnostický proces poskytnutím komplexného pohľadu na dynamiku červených krviniek. Využitím pokročilých techník strojového učenia výskum sľubuje výrazné zlepšenie presnosti krvnej diagnostiky, uľahčenie včasnej detekcie a liečby porúch.

Konečný cieľ práce je návrh a tréning modelu strojového učenia schopného klasifikovať červené krvinky s vysokou účinnosťou pomocou video údajov odvodených zo simulácií. To zahŕňa použitie stratégií hlbokého učenia, aby sa model naučil rozpoznávať zložité vzorce správania červených krviniek, čím sa položili základy pre vývoj robustného diagnostického rámca. Prostredníctvom tejto inovatívnej integrácie výpočtových

simulácií, vizuálnej analýzy a umelej inteligencie práca prispieva k pokroku v diagnostike krvi a otvára dvere pre personalizovanejšie a presnejšie riešenia zdravotnej starostlivosti pri liečbe porúch súvisiacich s krvou.

• Abstract

The thesis explores innovative methods for diagnosing blood-related disorders using a combination of machine learning techniques and simulations, particularly focusing on the analysis of red blood cells through the ESPResSo simulation module. This approach enables a deep dive into understanding red blood cell behavior under various physiological conditions by generating numerical outputs from simulations. The goal is to employ machine learning models, notably neural networks and random forests, to analyze these outputs for classifying RBCs based on changes in their elasticity, which can indicate the presence of blood disorders. This novel methodology seeks to bridge the gap between traditional diagnostic techniques and the vast potential of computational analysis, offering a more precise and efficient diagnostic tool.

The thesis's core goal is to extend beyond numerical data analysis by transforming the simulation outputs into dynamic visual representations. This involves generating videos to capture the behaviors of red blood cells in blood flow, further processed by machine learning models for classification purposes. This dual approach, encompassing both numerical and visual data, aims to enhance the diagnostic process by providing a comprehensive view of RBC dynamics. By leveraging advanced machine learning techniques, the research promises to significantly improve the accuracy of blood diagnostics, facilitating early detection and treatment of disorders.

The final goal of the thesis revolves around designing and training a machine learning model capable of classifying red blood cells with high efficiency, using video data derived from simulations. This involves employing deep learning strategies to teach the model to recognize complex patterns in red blood cell behavior, laying the groundwork for developing a robust diagnostic framework. Through this innovative integration of computational simulations, visual analysis, and artificial

intelligence, the thesis contributes to the advancement of blood diagnostics, opening the door for more personalized and precise healthcare solutions in treating blood-related disorders.

Abstrakt	4
Abstract.....	5
Úvod.....	7
Biologické zdôvodnenie analýzy krvi	8
Využitie strojového učenia v diagnostike krvi	10
Ciele práce	11
Stanovenie elasticity v priamom kanáli.....	12
Nastavenia simulačného vstupu	12
Predspracovanie a rozšírenie údajov	14
Testované typy architektúr neurónových sietí.....	15
Použitie NN na určenie elasticity RBC.....	16
Použitie regresnej neurónovej siete na klasifikáciu elasticity červených krviniek.....	17
Stanovenie elasticity v kanáli s prekážkami.....	17
Simulačné experimenty pre klasifikáciu červených krviniek .	18
Klasifikácia červených krviniek za použitia CNN.....	20
Pridanie fyzikálnej informácie	21
Klasifikácia od štyroch do dvoch tried.....	22
Použitie modelu náhodného lesa	23
Štatistická analýza elasticity na základe povrchu	26
Záver	28
Zoznam Publikovaných Prác.....	30

• Úvod

Mikrofluidika zahŕňa manipuláciu s tekutinami v mikrokanáloch a je kľúčová v rôznych vedeckých a lekárskech oblastiach. Tieto zariadenia, od jednoduchých ciest až po zložené komory, uľahčujú riadený tok tekutín a umožňujú presné biologické a chemické procesy. Mikrofluidná technológia výrazne ovplyvnila molekulárnu biológiu pokrokom v enzymatických analýzach, výskume DNA, chemickej syntéze a štúdiách proteínov. Okrem laboratórneho výskumu sú mikrofluidické zariadenia kľúčové v klinickej patológii, kde urýchľujú diagnostiku a hodnotenie závažnosti ochorenia. Tieto zariadenia sú nevyhnutné na monitorovanie rezistencie na antibiotiká hodnotením rýchlosti rastu mikroorganizmov. Navyše, ich prenosnosť ich robí neoceniteľnými pri diagnostike na mieste starostlivosti, čím zlepšuje prístup k zdravotnej starostlivosti v prostredí s obmedzenými zdrojmi tým, že umožňuje rýchle a decentralizované testovanie. Mikrofluidické technológie sú tiež v popredí monitorovania životného prostredia a objavovania liekov. Slúžia ako systémy včasného varovania pre toxíny vo vzduchu a vo vode a zefektívňujú procesy skríningu liekov, čo ich robí účinnejšie a cenovo dostupnejšie efektívnejšie. Tieto kontrolované mikroprostredia sú nápomocné pri štúdiu bunkových odpovedí na lieky, čím sa rozvíja personalizovaná medicína.

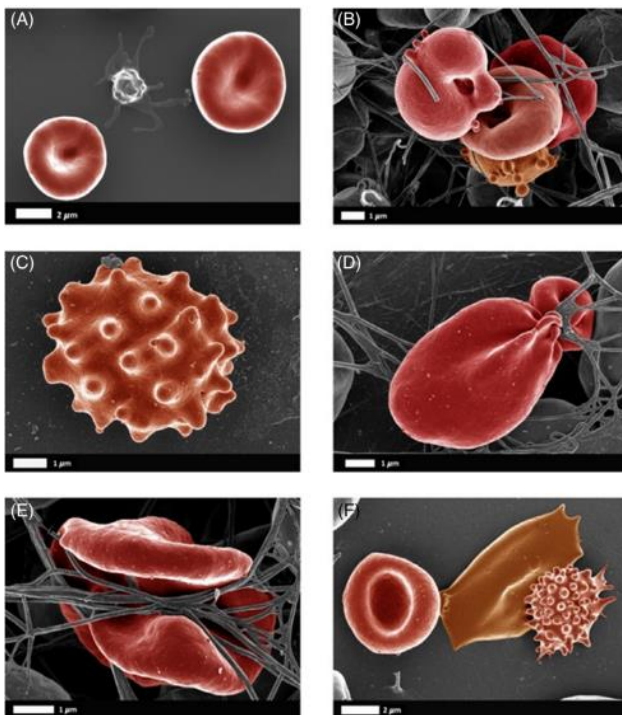
Výskumný tím v Cell in Fluid, Biomolecular Modeling & Computational Group (CIF) sa zameriava na využívanie mikrofluidických zariadení na zachytávanie a triedenie rakovinových buniek zo vzoriek krvi. Táto aplikácia je obzvlášť dôležitá pre včasnú diagnostiku rakoviny, pretože tieto zariadenia dokážu izolovať cirkulujúce nádorové bunky (CTC), ponúkajúce pohľad na metastatické ochorenia a zlepšujúce výsledky liečby. Prieskum CIF sa rozširuje na účinky mikrofluidného prostredia na červené krvinky, čo je kľúčové pre navrhovanie účinných a biokompatibilných systémov. Tento výskum pomáha nielen

pochopiť dynamiku buniek v rámci týchto zariadení, ale pomáha aj pri optimalizácii technológie pre širšie diagnostické aplikácie. Posun smerom k numerickým modelom, ktorý čelí výzvam fyzikálnych prototypov, ponúka pragmatické riešenie na simuláciu dynamiky tekutín v mikrofluidických systémoch. Tieto modely poskytujú nákladovo efektívnu alternatívu na pochopenie správania buniek v rôznych podmienkach, čím sa výrazne znižuje závislosť na nákladných a časovo náročných fyzických prototypoch.

So zložitou mikrofluidickými simuláciami generujúcich rozsiahle súbory údajov sa na spracovanie a analýzu týchto informácií čoraz viac používajú algoritmy strojového učenia. Tieto algoritmy zlepšujú simulácie a analýzu údajov, čo vedie k presnejším a spoľahlivejším diagnostickým nástrojom. Integrácia strojového učenia nielen optimalizuje funkčnosť zariadenia, ale otvára aj nové možnosti v mikrofluidickom výskume.

• **Biologické zdôvodnenie analýzy krvi**

Červené krvinky (RBC) sú rozhodujúce pre okysličenie tkanív a spoliehajú sa na ich deformovateľnosť, aby prešli mikrocirkuláciou. Zmeny v biomechanických vlastnostiach červených krviniek, najmä deformovateľnosť, sú spojené s rôznymi chorobami, vďaka čomu je ich štúdium nevyhnutné na pochopenie a diagnostiku týchto stavov. Tomaiuolo a kol. zdôrazňujú, že deformovateľnosť by mohla slúžiť ako spoľahlivý biomarker pre špecifické klinické stavy, čo podčiarkuje dôležitosť presného merania týchto vlastností. Tradičné metódy hodnotenia deformovateľnosti červených krviniek sú obmedzené nízkou priepustnosťou a statickými podmienkami, ktoré presne nereprezentujú fyziologické namáhanie, ktorým sú červené krvinky vystavené in vivo. Moderné techniky, ako je mikrokapilárny tok a mikrofluidika, ponúkajú dynamické, vysoko výkonné hodnotenia, ktoré simulujú skutočné mikrocirkulačné prostredia. Tieto pokroky umožňujú lepší pohľad na správanie sa červených krviniek za fyziologických podmienok a sú nápomocné pri diagnostike na mieste starostlivosti.



Obr. 1: Príklady eryptotických červených krviniek (RBC) pri zápale. (A) Zdravé červené krvinky s krvnými doštičkami; (B) diabetes typu 2; (C, D) Parkinsonova choroba; (E) reumatoidná artritída; (F) zdravá plná krv vystavená interleukínu-8 [1]

Dedičné poruchy: Poruchy ako sférocytóza a eliptocytóza ovplyvňujú membránu erytrocytov a sú významnými príčinami dedičných hemolytických anémie. Techniky, ako je ektocytometria a mikroskopia atómovej sily (AFM), boli kľúčové pri hodnotení zníženej deformovateľnosti v týchto podmienkach, čo ponúka hlbšie pochopenie základných cytoskeletálnych zmien.

Metabolické poruchy: Stavby ako diabetes (Obr. 1) a hypercholesterolémia ovplyvňujú deformovateľnosť červených krviniek v dôsledku zmeneného bunkového metabolizmu. Štúdie ukázali, že diabetes vedie k zvýšenej viskozite krvi a zmenám v architektúre membrán, ktoré sú zistiteľné pomocou pokročilých mikrofluidných zariadení, čo poukazuje na významné dôsledky pre mikrovaskulárne zdravie.

Oxidačný stres spochybňuje integritu červených krviniek tým, že ohromuje ich antioxidačné mechanizmy, čo vedie k poškodeniu, ktoré ovplyvňuje ich deformovateľnosť. Techniky, ako je mikrofluidná analýza, sa stali nevyhnutnými pri hodnotení toho, ako oxidačný stres ovplyvňuje deformovateľnosť červených krviniek, čo poskytuje vysoko výkonnú a nákladovo efektívnu metódu na hodnotenie tohto vplyvu v klinickom prostredí.

SCD ilustruje, ako môžu genetické mutácie ovplyvňujúce hemoglobín drasticky zmeniť tvar a funkciu červených krviniek, čo vedie k závažným klinickým symptómom, ako sú krízy bolesti a anémia. Ochorenie primárne ovplyvňuje schopnosť červených krviniek navigovať v kapilárach a efektívne dodávať kyslík. Prebiehajúci výskum v tejto oblasti sa zameriava na génové terapie zamerané na korekciu defektného génu zodpovedného za SCD, pričom ukazuje potenciál pokročilých genetických zásahov pri liečbe takýchto porúch.

Neustály pokrok v mikrofluidných technológiách a integrácia strojového učenia na analýzu údajov sľubuje zlepšenie diagnostických schopností pre choroby súvisiace s deformovateľnosťou červených krviniek. Tieto technológie pomáhajú nielen porozumieť základným vlastnostiam červených krviniek, ale aj pri vývoji cielených terapií pre stavy, ako je SCD a iné súvisiace poruchy.

• **Využitie strojového učenia v diagnostike krvi**

Neurónové siete (NN) spôsobili revolúciu v zdravotníctve tým, že zlepšili analýzu komplexných lekárskeho údajov, od diagnostických snímok až po komplexné záznamy pacientov. Tieto siete sú zdatné v identifikácii zložitých vzorov a anomálií, pomáhajú pri klasifikácii snímok, detekcii chorôb a predpovedaní rizík. Využitie NN v diagnostike krvi, konkrétne na klasifikáciu červených krviniek (RBC) prostredníctvom údajov z tokovej cytometrie a video sekvencií, predstavuje významný pokrok v tejto oblasti.

Číselné údaje z tokovej cytometrie: toková cytometria poskytuje podrobné informácie o červených krvinkách vrátane veľkosti, tvaru a optických vlastností. Konvolučné neurónové siete (CNN) spracúvajú tieto údaje, čo umožňuje presnú klasifikáciu podtypov RBC, čím sa zlepšuje pochopenie populácie RBC v rôznych zdravotných podmienkach.

Vizuálne sekvencie pre dynamickú analýzu červených krviniek: Okrem statických údajov zachytávajú videosekvencie dynamické správanie červených krviniek, ukazujúce ich pohyb a interakciu v krvnom obeh. Rekurentné neurónové siete (RNN) a siete s dlhou krátkodobou pamäťou (LSTM) analyzujú tieto sekvencie a učia sa časové vzorce nevyhnutné na diagnostikovanie chorôb, ako je malária a kosáčikovitá anémia.

CNN pre statické numerické údaje a RNN pre dynamické video informácie umožňuje komplexnú klasifikáciu červených krviniek. Táto metóda zvyšuje presnosť diagnostiky tým, že poskytuje poznatky založené na statických aj dynamických charakteristikách červených krviniek.

Zatiaľ čo aplikácia NN v klasifikácii RBC je sľubná, niekoľko výziev pretrváva. Medzi ne patrí potreba rozsiahleho predspracovania údajov, získavanie vysokokvalitných videosekvencií a vývoj rôznych súborov údajov na efektívne

učenie a overovanie. Prekonanie týchto prekážok je kľúčové pre pokrok v aplikácii NN v klinickom prostredí, čo potenciálne umožňuje analýzu v reálnom čase pre okamžitú diagnostickú spätnú väzbu.

Budúci výskum je zameraný na zlepšenie interpretovateľnosti výstupov NN s cieľom zvýšiť ich klinickú užitočnosť. Existuje tiež neustály tlak na vývoj schopností analýzy v reálnom čase v hematológii, ktoré by mohli transformovať diagnostiku RBC poskytnutím okamžitej klinickej spätnej väzby.

Integrácia NN s tradičnými a inovatívnymi diagnostickými metódami ponúka účinný nástroj na zvýšenie presnosti analýzy RBC v diagnostike krvi.

○ **Ciele práce**

Hlavné ciele tejto diplomovej práce sú zamerané na zlepšenie diagnostických schopností pre choroby súvisiace s krvou prostredníctvom inovatívnej integrácie strojového učenia a výpočtových simulácií, špecificky zameraných na dynamické správanie a vlastnosti červených krviniek (RBC).

Hlavným cieľom tohto výskumu je vyvinúť robustný diagnostický rámec, ktorý využíva simulačný nástroj ESPResSo na modelovanie komplexnej dynamiky červených krviniek v krvnom obeh. Využitím sofistikovaných výpočtových simulácií práca generuje numerické výstupy, ktoré poskytujú pohľad na správanie červených krviniek za rôznych fyziologických podmienok. Tieto výstupy slúžia ako základ pre použitie pokročilých techník strojového učenia – vrátane neurónových sietí a náhodných lesov – na analýzu a klasifikáciu červených krviniek na základe zmien v ich elasticite. Takéto zmeny svedčia o rôznych krvných poruchách, vďaka čomu je tento prístup kľúčový pre včasnú detekciu a liečbu.

Práca navrhuje duálny prístup kombináciou numerickej a vizuálnej analýzy dát. Spočiatku sa zameriava na numerické údaje

získané zo simulácií s cieľom vyvinúť modely strojového učenia, ktoré dokážu klasifikovať červené krvinky na základe pozorovaných zmien. Táto fáza zahŕňa stratégiu hlbokého učenia na rozpoznanie komplexných vzorcov v správaní červených krviniek a poučenie sa z nich. Následne sa výskum rozširuje na transformáciu týchto simulačných výstupov do dynamických vizuálnych reprezentácií, ako sú videá, ktoré sa potom používajú na tréning modelov strojového učenia pre ďalšie klasifikačné úlohy. Cieľom tejto metodiky je poskytnúť komplexný pohľad na dynamiku červených krviniek, čím sa zvýši presnosť a účinnosť diagnostiky krvi.

Inovatívny aspekt tejto práce spočíva v integrácii výpočtových simulácií so strojovým učením. Vytvorením simulácií, ktoré presne odrážajú fyziologické podmienky toku krvi, výskum poskytuje bohatý súbor údajov pre tréning modelov strojového učenia. Tieto modely sú navrhnuté tak, aby pochopili a predpovedali správanie a charakteristiky červených krviniek za rôznych podmienok, čím posúvajú hranice tradičných diagnostických metód, ktoré sú často obmedzené pomalšími a invazívnymi technikami.

Očakáva sa, že tento výskum výrazne zlepši presnosť diagnostiky krvi. Tým, že poskytuje metódu na rýchle a presné zisťovanie a klasifikáciu porúch súvisiacich s krvou, podporuje stratégiu včasnej intervencie, ktoré sú kľúčové pre efektívne výsledky liečby. Okrem toho, metodológie vyvinuté v tejto práci majú potenciálne aplikácie nad rámec diagnostiky krvi, vrátane iných oblastí zdravotnej starostlivosti, kde dynamika tekutín zohráva kľúčovú úlohu.

• Stanovenie elasticity v priamom kanáli

Výhoda výpočtových simulačných modelov spočíva v prístupe ku komplexným informáciám o pohybe červených krviniek v krvnom riečisku, ktoré prevyšujú to, čo je možné zaznamenať pri experimentoch na fyzikálnych zariadeniach. Červené krvinky, ktoré sú elastickými entitami v krvnej plazme a tvoria prevládajúcu zložku krvi, sú kľúčové pre presné simulácie toku krvi. Využívame údaje o polohe povrchových bodov na červených krvinkách v priamom kanáli ako vstupný súbor údajov pre tréning neurónových sietí. Uvádzame výsledky využitia neurónových sietí na posúdenie elasticity červených krviniek v krvnom obehú z numerických výstupov simulačného modelu.

Snaha určiť elasticitu červených krviniek pomocou simulačných experimentov bola predmetom výskumu [2]. Ich primárnym cieľom bola kategorizácia elasticity RBC do vopred definovaných tried. V tejto kapitole rozširujeme túto líniu skúmania a premieňame ju na regresnú úlohu. Tu je naším cieľom určiť hodnotu regresnej funkcie charakterizujúcej elasticitu každého jednotlivého RBC. Našou konečnou, aj keď vzdialenou ambíciou je zistiť elastické vlastnosti alebo elastické koeficienty červených krviniek z videozáznamov laboratórnych experimentov.

Naš inovatívny prístup využíva údaje odvodené z počítačových simulácií založených na laboratórnych experimentoch. Tieto údaje ponúkajú výnimočnú presnosť, vďaka čomu sú vhodné na učenie pod dohľadom. Významnou výhodou je, že tento súbor údajov neguje potrebu manuálnych anotácií, pretože počítač môže túto úlohu zvládnuť autonómne. Využitím simulácií môžeme využiť dáta vhodné na vstup do neurónových sietí. Aj keď máme možnosť použiť výlučne údaje získané z videoanalýz týchto experimentov, rozhodli sme sa zahrnúť aj komplexné simulačné údaje pre metodologické porovnania.

Cieľom je vyhodnotiť kvalitu odhadov koeficientov elasticity získaných z obmedzených údajov v porovnaní s tými, ktoré boli získané pomocou celého simulačného súboru údajov.

○ ***Nastavenia simulačného vstupu***

Súčasný výskum skúma aplikáciu fyzicky informovanej neurónovej siete, ako bolo zdôraznené v predchádzajúcich prácach [3][4]. V našej práci využívame implicitné informácie modelov strojového učenia, čím zabezpečujeme súlad s fyzikálnymi princípmi.

Zdrojové údaje pre odhad k_s koeficientu prostredníctvom strojového učenia pochádzajú z niekoľkých simulačných experimentov. Tieto simulácie sú dôsledne zosúladené so skutočnými experimentmi a boli predtým použité v štúdií [5].

V týchto simuláciách sa použil štandardizovaný kanál charakterizovaný kvádrovým tvarom so štyrmi stenami s rozmermi $60 \times 40 \times 40 \mu\text{m}$. Periodické vlastnosti toku kvapaliny boli stanovené v smere osi x . Tekutina bola diskretizovaná do trojrozmernej mriežky s priestorovým rozlíšením $1 \mu\text{m}$. Kvapalina mala kinematickú viskozitu ($1.3 \times 10^{-6} \text{ m}^2/\text{s}$) a hustotou ($1.025 \times 10^3 \text{ kg/m}^3$). Interakcia medzi tekutinou a predmetmi bola udržiavaná prostredníctvom koeficientu trenia $1,414$ a vonkajšími silami indukovaného toku na hodnoty zaručujúce maximálnu rýchlosť približne $0,03 \text{ m/s}$.

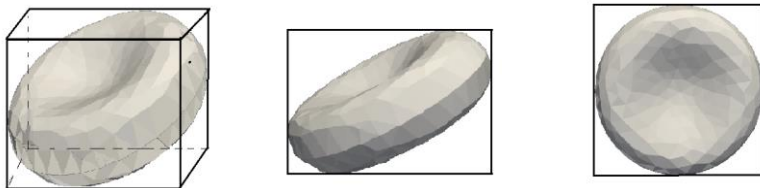
Interakcia medzi červenými krvinkami, známa ako *interakcia bunka-bunka*, bola modelovaná pomocou potenciálu *membrane_collision* charakterizovaného parametrami $mc_a = 0,01$, $mc_n = 1,0$ a $mc_{cut} = 0,4$. Interakcie medzi bunkami a stenami kanála boli reprezentované potenciálom *soft_sphere* definovaným parametrami $soft_a = 0,00035$, $soft_n = 1,0$ a $soft_{cut} = 0,5$.

Geometria červených krviniek bola zachytená pomocou povrchovej sieťky obsahujúcej 374 uzlov. Vo svojom uvoľnenom stave tieto bunky nadobudli charakteristický bikonkávny tvar s

rozmermi $7,82\text{ }\mu\text{m}$ x $7,82\text{ }\mu\text{m}$ x $2,58\text{ }\mu\text{m}$ a objemom $90,75\text{ }\mu\text{m}^3$. Pozoruhodné je, že bunky boli naplnené rovnakou tekutinou ako v ich bezprostrednom okolí.

Počiatkové usporiadanie krvných doštičiek bolo pri simuláciách náhodné a počet zaznamenaných krokov simulácie bol 3400 pre Sim3a, Sim3b a Sim3c a 1240 krokov pre Sim9a a Sim9b, čo zodpovedalo Pohyb RBC v kanáli asi $5,5\text{ mm}$. Zaznamenané kroky zodpovedali každých 2000 krokov simulácie, čo je vnútorný krok simulácie dostatočne podrobný na nácvik metódy. Z dôvodu potreby stabilizácie toku v nábehovej časti experimentu nebolo použitých prvých 300 záznamov výstupov zo simulácie. Na vytvorenie vyváženého súboru údajov bol každý typ červených krviniek kategorizovaný podľa elasticity reprezentovaný rovnakým množstvom údajov. V experimentoch Sim3 bolo simulovaných 9 typov elasticity, každý so 6 červenými krvinkami a každý typ mal 3100 záznamov. Pre Sim9 bolo tiež 9 rôznych typov, so 6 krvinkami každého typu a 940 záznamami. Celkovo bolo simulovaných 54 červených krviniek v oboch prípadoch, Sim3 a Sim9. Podobný postup bol použitý v štúdiu [2].

V každom internom simulačnom kroku modulu ESPreSso sa vypočíta aktuálna poloha každého nestacionárneho bodu v toku krvi, a teda všetkých triangulačných bodov červených krviniek. Vzhľadom na obrovský rozsah týchto údajov boli zvyčajne uložené základné informácie o polohe a rýchlosti životne dôležitých bodov každej krvinky, vrátane čísla kroku (cyklu) simulácie, súradníc stredu simulovanej bunky $[x, y, z]$, rýchlosť stredu bunky určená jej zložkami v smere osí x, y, z , súradnice x, y, z predstavujúce krajné body triangulácie bunky (podľa na minimálne a maximálne súradnice pozdĺž každej osi (ako je znázornené na obrázku nižšie) a rýchlosti extrémnych bodov bunky určené jej komponentmi v smere osí x, y, z , ako aj objem alebo povrch bunky.



Pri výbere a úprave údajov sme zvažovali, aké informácie môžeme získať zo skutočných videozáznamov. Zatiaľ čo informácie obsiahnuté vo výstupoch zo simulácie nám umožňujú určiť krajné body červenej krvinky v reálnom experimente, ostatné parametre ako stred bunky, rýchlosť jej pohybu, pohyb krajných bodov, objem alebo celkový povrch bunky je oveľa náročnejšie získať, ak vôbec. Preto sme sa v našej práci zamerali na využitie dát, ktoré predstavujú premietanie informácií z 3D do dvojrozmerného priestoru pozdĺž osi xy a xz . Odčítaním polôh krajných bodov v smere každej osi by sme tiež mohli odvodiť informácie o veľkosti obdĺžnikového "ohraničujúceho rámčeka" (alebo kocky), ktorý obklopuje krvinku v každom časovom kroku.

○ **Predspracovanie a rozšírenie údajov**

Súbor údajov získaný zo simulačných experimentov potreboval transformáciu, aby bol vhodný na tréovanie modelu strojového učenia. Vzhľadom na to, že simulačné dáta majú za cieľ nahradiť skutočnú video informáciu, informácie o rýchlosti bunky sú implicitne zahrnuté na základe zmeny hodnôt súradníc x v po sebe idúcich záznamoch. Výsledný súbor údajov bol teda rozdelený do časových okien, čo zodpovedá sekvencii po sebe idúcich simulačných záznamov pri spracovaní videa. Rozhodujúcim hyperparametrom modelu sa stáva dĺžka sekvencie, reprezentovaná veľkosťou okna w .

Vo výpočtových experimentoch zahŕňajúcich tréovanie jednotlivých neurónových sietí boli použité veľkosti okien z

množiny $w = 5, 10, 20, 30, 40, 50$. Po výbere w sa celý súbor údajov rozdelil na časové sekvencie veľkosti w , výsledkom čoho boli tréningové údaje (x_{wi}, y_i) , kde x_{wi} predstavuje časovú sekvenciu, a y_i je skutočná hodnota koeficientu pružnosti k_s pre túto sekvenciu.

Na urýchlienie procesu trénovania modelu sme použili šandardizačné a normalizačné techniky. Šandardizácia bola vykonaná na hodnotách krajných bodov pozdĺž osí y a z a rozmerov ohraničujúceho obdĺžnika (alebo kocky). Táto transformácia zabezpečila, že dáta mali priemer 0 a smerodajnú odchýlku 1. Atribúty sme potom upravovali samostatne pre každý tréningový príklad obsahujúci informácie o súradniciach krajných bodov pozdĺž osi x . Pre každý záznam v rámci časového okna sme odrátali minimálnu hodnotu atribútu x_i od atribútu x_i . Tento prístup nám umožnil normalizovať príklady tréningu a zároveň implicitne zachovať informácie o rýchlosti pohybu červených krviniek.

Vytvorené údaje boli následne rozdelené do troch častí (tréning, validácia, testovanie. Toto rozdelenie bolo vykonané, aby sa zabránilo úniku údajov. S cieľom rozšíriť tréningové údaje pre strojové učenie boli pôvodné pozície transformované pridaním šumu. Šum bol generovaný z náhodného normálneho rozdelenia s priemerom 0 a štandardnou odchýlkou 1 a bol vynásobený konštantou 0,1 pre každú zložku príkladu tréningu. Navyše, náhodný posun (s) bol vygenerovaný z rovnomerného rozdelenia v rámci intervalu $(-0,25, 0,25)$ pre každý príklad tréningu. Počet rozšírení (a) bol určený nasledovne:

$$a = 10000 / ((3100 - 50)/(w)) - 1$$

a potom sa a zaokrúhli na celé číslo. Množstvo tréningových dát sa tak zvýšilo na približne 380 000 príkladov.

○ **Testované typy architektúr neurónových sietí**

Architektúra LSTM (Long-Short Term Memory) citeyu2019review sa zvyčajne používa na vstupné dáta vo forme časových sekvencií. Naša sieť využívajúca túto architektúru pozostávala zo štyroch vrstiev s 512, 64, 32 a 10 skrytými neurónmi. Využila hyperbolickú tangentovú aktivačnú funkciu a rekurentnú sigmoidnú aktivačnú funkciu. Po každej vrstve LSTM nasledovala vrstva odpadnutia, v ktorej bolo vyradených 10% neurónov. Výstup bol potom sploštený (`keras.layers.Flatten()`) a prešiel cez pár plne prepojených vrstiev, jedna s 1024 neurónmi, druhá s 512 neurónmi, obe s použitím funkcie aktivácie ReLU. Nakoniec plne prepojená výstupná vrstva s jedným neurónom a lineárnou aktivačnou funkciou slúžila ako výstupná vrstva.

Architektúra CNN-LSTM kombinuje konvolučné vrstvy CNN na extrakciu prvkov zo vstupných údajov s LSTM na predikciu sekvencie. Kombinácia týchto vrstiev je motivovaná výskumom a štúdiami ako [6], ktoré naznačujú zlepšený výkon LSTM s touto architektúrou. Je vhodný pre problémy, ktoré zahŕňajú dáta s priestorovými štruktúrami (napr. obrazové pixely) alebo časovo štruktúrovanými vstupmi/výstupmi (napr. video snímky alebo text).

Implementovali sme dve verzie siete CNN-LSTM. Prvá, CNN-LSTM Conv1D, používala 1D konvolúciu pri prechádzaní časovou sekvenciou. Druhá, CNN-LSTM Conv2D, používala 2D konvolúciu. Aby sme znížili rozptyl vstupných údajov, vstup prešiel cez dve konvolučné vrstvy s 256 filtrami, veľkosť kroku 1 a aktiváciou ReLU. Pre CNN-LSTM Conv1D bola veľkosť filtra pre obe vrstvy trojnásobkom šírky časového okna. Pre CNN-LSTM Conv2D bola veľkosť filtra trojnásobkom šírky časového okna pre prvú vrstvu a 4x3 pre druhú vrstvu. Po konvolučných vrstvách bol výsledok upravený spojením s veľkosťou filtra 2x2. Multivariačný medzivýsledok bol sploštený (`keras.layers.Flatten()`) a odovzdaný vrstve LSTM s 256 skrytými

neurónmi a funkciou aktivácie ReLU. Nakoniec bol výstup z vrstvy LSTM pripojený k plne prepojenej vrstve s 512 neurónmi pomocou aktivácie ReLU a sieť bola ukončená lineárnou výstupnou vrstvou obsahujúcou jeden neurón.

Všetky experimenty sa uskutočňovali pomocou Pythonu 3.8 s knižnicou Tensorflow (Keras) použitou na konštrukciu neurónových sietí. učenie prebiehalo na počítači s procesorom AMD Ryzen 55600H, grafickou kartou Radeon, 16 GB RAM a grafickou kartou NVIDIA GeForce RTX 3060 Laptop.

○ **Použitie NN na určenie elasticity RBC**

Vo výskume sme trénovali 84 rôznych modelov neurónových sietí na analýzu elasticity červených krviniek (RBC) s použitím rôznych architektúr, súborov údajov a veľkostí časového okna w . Použili sme strednú absolútnu percentuálnu chybu (MAPE) ako stratovú funkciu, pričom hodnoty MAPE naznačujú, že so zvyšujúcou sa veľkosťou tréningového okna rastie aj MAPE, čo môže byť výsledkom šumu a skreslenia v dátach.

Najnižšie hodnoty MAPE, ktoré indikujú vyššiu presnosť modelov, boli dosiahnuté pomocou údajov zo všetkých troch osí xyz , kde sa chyba znížila na 5%. Tieto výsledky potvrdzujú hypotézu, že informácie o elasticite RBC možno efektívne získať z použitých dát. Najpresnejší model, CNN-LSTM Conv2D, bol aplikovaný na údaje pozdĺž osí xy , xyz , a xy_xz s veľkosťami okien w postupne 3, 3, a 5. Pre osi xz sa najlepšie osvedčil LSTM s $w=5$, čo vyústilo v MAPE hodnotu 20,46%. Kombinácia dvoch 2D projekcií pomocou CNN-LSTM Conv2D s $w=5$ dosiahla MAPE 7,97%.

Distribúcia MAPE pre rôzne typy elasticity RBC bola vizualizovaná na grafických znázorneniach. Napríklad v modeli CNN-LSTM Conv2D s $w=3$ pre dáta xyz sa najväčšie stredné hodnoty MAPE vyskytli pri elasticite 0,03 a 0,1, čo ilustrujú grafy pre rôzne súbory údajov. Tieto grafy ukazujú variabilitu

predpovedaných hodnôt k_s v závislosti od použitého modelu a nastavení. Takéto podrobné analýzy sú kľúčové pre pochopenie a optimalizáciu výkonnosti modelov pri klasifikácii a predikcii vlastností červených krviniek.

○ **Použitie regresnej neurónovej siete na klasifikáciu elasticity červených krviniek**

V tejto časti transformujeme výstupy z regresného modelu na jednoduchý klasifikátor, pričom využívame rovnaké simulačné dáta ako v predchádzajúcej sekcii. Používame dva prístupy: tréning klasifikačnej neurónovej siete a priamy výstup z regresnej neurónovej siete, kde sa priemerné hodnoty elasticity priradzujú do príslušných kategórií na základe stanovených prahov.

Architektúra klasifikačnej neurónovej siete sa líši od regresnej iba v poslednej vrstve, kde jeden výstupný neurón je nahradený deviatimi na reprezentáciu deviatich kategórií, s použitím softmax funkcie namiesto lineárnej. Funkcia strát je zmenená na kategorickú krížovú entropiu a cieľové premenné sú prevedené na one-hot kódovanie. Táto metóda ukázala lepšie výsledky pri priamom trénovaní pre klasifikačné úlohy, s výrazným zlepšením až o 11.75% pre modely trénované na projekciách do 2D pozdĺž osi xz.

Výsledky validácie na rôznych simulačných dátach odhalili značné zhoršenie výkonnosti modelu napriek snahám o redukcii preučenia pridaním šumu a dropout vrstiev. Toto pokles môže byť spôsobené dátami obsahujúcimi informácie o osiach y a z. Model mohol byť prečuný sústredením sa na tieto časti údajov a predpovedaním koeficientu elasticity na základe polohy bunky v kanáli. Tento problém naznačuje, že CNN s architektúrou CNN_LSTM_Conv2D je výkonná, čo potvrdzujú lepšie výsledky vo všetkých experimentoch s použitím tejto štruktúry.

Kód je dostupný na https://github.com/molcan23/RBC_NN.

Confusion matrix W_3_A_9_X_xyz (in %)

0.005	99.12	0.00	0.88	0.00	0.00	0.00	0.00	0.00	0.00
0.009	0.00	96.78	0.00	2.88	0.00	0.34	0.00	0.00	0.00
0.015	0.57	0.00	99.43	0.00	0.00	0.00	0.00	0.00	0.00

• **Stanovenie elasticity v kanáli s prekážkami**

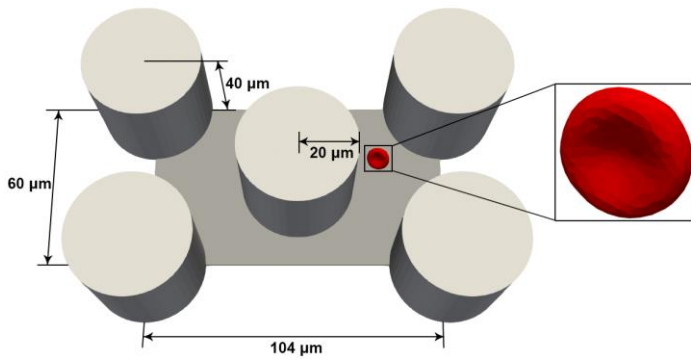
Táto kapitola skúma geometrické charakteristiky červených krviniek na videozáznamoch alebo číselných údajoch z kanála s prekážkami. Simulačný model toku krvi ponúka nové príležitosti na analýzu toku krvi a jeho vlastností prostredníctvom NN. S presným simulačným modelom tento výskum skúma presnosť identifikácie vlastností buniek z 2D obrazu simulácie v porovnaní s úplnými údajmi modelu. Zameriava sa tu nielen na replikáciu skutočného experimentu, ale aj na skúmanie schopností NN pri kategorizácii buniek s rôznou elasticitou, keď sú poskytnuté úplné informácie (presné polohy častíc vo všetkých troch rozmeroch) oproti iba dvojrozmerným video dátam. Použitie CNN na identifikáciu elasticity pohyblivých krvných buniek predstavuje druhý pokus o riešenie tohto problému.

CNN, inšpirované biologickými systémami [7], sú široko používané v hlbokom učení pre úlohy klasifikácie obrázkov a videa. Primárna výhoda CNN spočíva v ich schopnosti analyzovať obsah obrazu a zároveň explicitne využívať priestorovú štruktúru prostredníctvom funkcií, ako sú lokálne filtre, konvolúcia a maximálne združovanie.

○ ***Simulačné experimenty pre klasifikáciu červených krviniek***

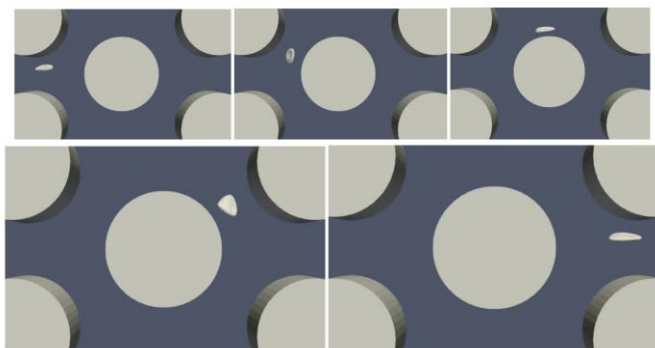
Zdroj údajov na klasifikáciu zdravia červených krviniek bol získaný prostredníctvom viacerých simulačných experimentov. Všetky simulácie sa uskutočňovali pri konzistentných parametroch kanála a toku tekutiny. Kanál mal kvádrový tvar s rozmermi $104 \times 60 \times 40 \mu\text{m}$ a tekutina bola diskretizovaná do trojrozmernej siete s priestorovým krokom $1 \mu\text{m}$. RBC pretekajú priestorom s prekážkami. Simulovaný kanál obsahoval päť valcov pôsobiacich ako prekážky, ktoré obmedzovali oblasť toku krvi a

vyvolávali elasticitu červených krviniek, ako je znázornené na obrázku:



Do úvahy sa brali štyri úrovne elasticity červených krviniek. Zdravé červené krvinky s koeficientom tuhosti (k_s) 0,005 a najmenej elasticke červené krvinky predstavujúce bunky infikované maláriou v štádiu 3 choroby s hodnotou $k_s = 0,03$. Zvyšné dve úrovne elasticity RBC boli rovnomerne rozdelené medzi zdravé a maláriou infikované RBC s hodnotami $k_s = 0,0133$ a $0,0216$.

Každá simulácia v našom výskume zahŕňala 36 červených krviniek, pričom deväť buniek predstavovalo každú úroveň elasticity. Simulovaný kanál má periodickú štruktúru, čo znamená, že keď RBC opustí kanál, znova sa objaví na začiatku. Odstraňujeme počiatočné a konečné priechody červených krviniek z dôvodu ich neúplnosti. Výsledkom je, že každá RBC dokončí približne 20--21 prechodov cez simulovaný kanál:



Vzhľadom na malý počet tréningových príkladov máme dve možnosti, ako toto obmedzenie prekonať. Prvá možnosť zahŕňa využitie techník zväčšenia údajov na rozšírenie našej tréningovej sady a zväčšenie jej veľkosti. Rozšírenie údajov zahŕňa použitie rôznych transformácií na existujúce príklady, aby sa vytvorili nové variácie. V našom prípade môžeme vykonávať vertikálne otočenia (pre zachovanie smeru toku krvi) a rotácie (pri zohľadnení zachovania smeru toku krvi).

Druhou možnosťou je využiť predtrénované modely, ktoré nevyžadujú učenie od nuly, a preto vyžadujú menej tréningových príkladov. Vybrali sme dva predtrénované modely, EfficientNet v2 B0 a ResNet50, ktoré preukázali úspešné výsledky v rôznych obrazových a video úlohách. Tieto modely vyžadujú menší počet tréningových príkladov na doladenie posledných vrstiev NN, vďaka čomu sú vhodné pre náš malý tréningový súbor údajov.

Na prípravu súboru údajov sme zo simulačných údajov vygenerovali videozáznamy jednotlivých červených krviniek. Tieto videá sa používajú na účely učenia a overovania. Keďže simulácie vykonané v ESPResSo poskytujú trojrozmerné informácie o toku simulovaných červených krviniek, premietli sme tieto údaje do dvojrozsmernej roviny. Dosiahli sme to vytvorením 2D videa, ktoré zachytáva šírku a dĺžku kanála bez ohľadu na rozdiely v hĺbke. Táto transformácia nám umožnila efektívne analyzovať a klasifikovať správanie RBC vo videách pomocou techník klasifikácie obrázkov a videí.

○ ***Klasifikácia červených krviniek za použitia CNN***

Naša sieť využívala ukážky videa v tvare $N \times T \times H \times W \times C$, kde N predstavuje veľkosť dávky (batchu), T je počet snímok vo videu, H je výška, W je šírka a C je počet kanálov, čo je 3 (predstavuje červenú, zelenú a modrú). Následne sme video zmenili na čiernobiely formát, čím sme znížili počet kanálov z 3 na 1. Na klasifikáciu videa sme použili časovo rozloženú vrstvu, ktorá umožňuje modelu klasifikovať videozáznamy. Optimalizátorom siete môže byť Adam alebo SGD.

Prebehla rozsiahla optimalizácia hyperparametrov. Najvyššia presnosť sa dosahuje s modelom EfficientNet v2 B0 pomocou

Hyperparameter	Value
Dropout	0.1
Regularizer	L1 + L2
Initial learning rate	1×10^{-5}
Final learning rate	1×10^{-5}
Decay steps	1×10^4

optimalizátora Adam s najlepšími hyperparametrami:.

Po analýze misklasifikácie pre validačný súbor sme zistili, že primárna výzva v klasifikácii spočíva v rozlíšení medzi RBC so zníženou elasticitou, zatiaľ čo rozdiel medzi zdravými a chorými RBC je viac presný. Následne sme sa rozhodli trénovať NN na binárnu klasifikáciu. V tomto nastavení prvá trieda zahŕňa zdravé červené krvinky ($k_s = 0,005$) a druhá trieda zahŕňa všetky bunky so zníženou elasticitou. Vykonalí sme optimalizáciu hyperparametrov pre dva typy NN s dvoma rôznymi možnosťami optimalizácie.

Prechod zo 4-triedy na 2-triedny problém klasifikácie viedol k zvýšeniu konečnej presnosti, aj keď nie tak výrazne.

Experimentovali sme aj s váženou klasifikáciou pomocou kombinácie EfficientNet v2 B0 a optimalizátora Adam, čo viedlo k marginálnemu zvýšeniu presnosti na 61,74%, čo znamená zlepšenie len o 0,02%.

Ďalším prístupom bolo oddelenie zdravých a chorých červených krviniek po ich zaradení do 4 tried, kde trieda 0 predstavovala zdravé červené krvinky, zatiaľ čo triedy 1, 2 a 3 predstavovali choré červené krvinky. S týmto prístupom sme dosiahli presnosť klasifikácie 93,54 %

■ Pridanie fyzikálnej informácie

Aby sme zvýšili výkon siete, pridali sme informácie o základnej fyzike, konkrétne o rýchlosti tekutiny prúdiacej v kanáli. Naša hypotéza bola, že pridaním fyzikálnej informácie by sa NN mohla naučiť klasifikovať RBC lepšie ako bez nej.

Fyzikálnu informáciu sme pridali vytvorením novej vetvy NN. Informácia prechádza cez vrstvu Rescale, tri 2D konvolučné vrstvy s množstvom filtrov 32, 16, 8, po ktorých nasleduje maximálna združovacia vrstva, vyrovnávacia vrstva a hustá vrstva s 1024 skrytými jednotkami. Výsledok je zreťazený s predposledným výstupom hlavnej vetvy pôvodnej NN, ktorý je odovzdaný poslednej plne prepojenej vrstve.

Vo všeobecnosti nie je efektívne pridávať rovnaké informácie ku každému príkladu učenia pre NN. Dôvodom je, že NN sa učia vzorce a vzťahy v údajoch prostredníctvom variácií a rozdielov medzi príkladmi. Keď všetky príklady obsahujú rovnaké informácie, sieť medzi nimi nedokáže rozlíšiť a nemusí sa naučiť relevantné vzory, ktoré sú potrebné na presné predpovede. Môžu však nastať prípady, keď pridanie rovnakých informácií ku každému príkladu učenia môže byť užitočné. Napríklad, ak pridané informácie poskytujú nejaké kontextové alebo základné informácie, ktoré sú relevantné pre všetky príklady, môžu pomôcť sieti učiť sa efektívnejšie. Vo všeobecnosti je dôležité starostlivo zvážiť informácie, ktoré sa pridávajú ku každému príkladu učenia a ako môžu ovplyvniť schopnosť siete učiť sa a zovšeobecňovať.

Najvýkonnejší model je EfficientNet v2 B0 s optimalizátorom Adam, pre ktorý bola výsledná presnosť 91,48%. Tá sa učila klasifikovať 4 triedy krviniek a následne v čase validácie rozdeľovala krvinky len do dvoch tried (zdravé a choré). Získané výsledky odhaľujú niekoľko kľúčových pozorovaní týkajúcich sa výkonnosti rôznych klasifikačných prístupov. Po prvé, pri porovnaní dvojtriednej klasifikácie (či už vázenej alebo nevázenej) s prevodom štvortriednej klasifikácie na dvojtriednu je

zrejmé, že druhá trieda dosahuje výrazne lepšie výsledky. To znamená, že dodatočné informácie prítomné v klasifikácii štyroch tried prispievajú k zvýšeniu celkovej presnosti modelu.

Okrem toho sa zdá, že začlenenie informácií súvisiacich s fyzikou vo forme tepelnej mapy je menej efektívne v porovnaní s použitím vektorov rýchlosti ako vstupu. To naznačuje, že vektory rýchlosti nesú zmysluplnejšie a diskriminatívnejšie informácie na presnú klasifikáciu. Použitie vektorov rýchlosti ako vstupu pravdepodobne umožňuje modelu zachytiť dynamické vzory a lepšie pochopiť pohybové charakteristiky analyzovaných údajov.

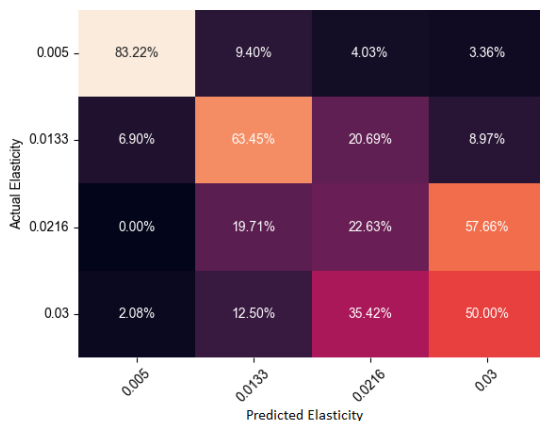
Je však zaujímavé poznamenať, že ani po zahrnutí informácií o vektore rýchlosti nedochádza k podstatnému zlepšeniu konečnej presnosti siete. Konkrétne, NN, ktorá sa spolieha výlučne na poskytnutý vstup bez akýchkoľvek dodatočných fyzických informácií, dosahuje presnosť 93,56%. Naproti tomu sieť využívajúca rýchlostné vektory dosahuje o niečo nižšiu presnosť 91,48 %. To naznačuje, že hoci zahrnutie vektorov rýchlosti môže ponúknuť cenné informácie, nemusí sa nevyhnutne premietnuť do významného zlepšenia výkonu klasifikácie.

○ ***Klasifikácia zo štyroch do dvoch tried***

Získané výsledky odhaľujú významné poznatky o výkonnosti rôznych klasifikačných prístupov. Pozoruhodný postreh spočíva v porovnaní medzi dvojtriednou klasifikáciou a štvortriednou klasifikáciou prevedenou na dvojtriednu klasifikáciu. Dvojtriedna klasifikácia sa týka klasifikačnej úlohy, ktorá rozlišuje medzi dvoma špecifickými triedami (napr. zdravý a chorý), zatiaľ čo

štvortriedna klasifikácia prevedená na dvojtriednu zahŕňa zlúčenie viacerých tried do dvoch širších kategórií.

Výsledky ukazujú, že štvortriedna klasifikácia prevedená na dvojtriednu poskytuje výrazne lepší výkon v porovnaní s



dvojtriednou klasifikáciou. To naznačuje, že zahrnutie ďalších tried do tréningového procesu poskytuje cenné informácie, ktoré zlepšujú celkovú presnosť modelu. Trénovanie modelu na množine údajov obsahujúcej viacero tried umožňuje zachytiť širšie spektrum vzorov, odchýlok a charakteristík údajov. Dôležitý je aj premyslený návrh klasifikačných úloh a výber vhodnej reprezentácie triedy na dosiahnutie optimálneho výkonu. Využitím dodatočných informácií, ktoré poskytuje štvortriedna klasifikácia, model získava hlbšie pochopenie základných údajov, čo vedie k vyššej presnosti pri rozlišovaní medzi širšími kategóriami. To podčiarkuje dôležitosť zváženia relevantnosti a zahrnutia dodatočných informácií do klasifikačných úloh.

Významný je výber vhodných vstupov v klasifikačných modeloch. V tomto prípade sa začlenenie vektorov rýchlosti ako vstupných znakov ukazuje ako efektívnejšie v porovnaní s použitím reprezentácie tepelnej mapy. Architektúra siete a

optimalizačné techniky sú hlavnými faktormi ovplyvňujúcimi konečný výkon. Výsledky zdôrazňujú dôležitosť starostlivého zvažovania rôznych faktorov, vrátane návrhu klasifikačnej úlohy, reprezentácie triedy, výberu vstupných funkcií, architektúry siete a optimalizačných techník, aby bolo možné dosiahnuť s optimálnou presnosťou klasifikácie.

Kód je dostupný na https://github.com/molcan23/rbc_video_classification.

○ **Použitie modelu náhodného lesa**

Na výpočet prediktorov zo simulácie používame trianguláciu bunky. Výstupom simulácie je pre každú bunku poloha každého z 374 uzlov triangulácie, určená tromi súradnicami v trojrozmernom priestore. Z týchto dát možno vypočítať rôzne charakteristiky, ktoré následne používame ako prediktory pri klasifikácii. Celkom sme vytvorili 41 prediktorov. Tie sme podľa odhadovanej zložitosti ich získania z reálneho experimentu rozdelili do niekoľkých sád. Na základe toho sme vytvorili 6 testov odstupňovaných podľa množstva prediktorov, ktoré pri nich používame: V prvom teste používame len najľahšie získateľné prediktory, v poslednom teste všetky.

- 1. sadu tvoria len dva prediktory: rozmery obdĺžnika, v ktorom sa sledovaná bunka nachádza, t. j. rozmer bunky v smere x-ovej a v smere y-ovej osi. Tieto hodnoty možno získať jednoducho zo (statickej) snímky bunky.
- V 2. sade sú navyše rýchlosti a zmeny rýchlosti bunky v x-ovom a y-ovom smere. Tieto údaje možno dopočítať z viacerých po sebe idúcich snímok.
- 3. sada obsahuje navyše aj rozmer a rýchlosti v smere z-ovej osi. Tieto hodnoty možno získať zo snímok vytvorených z iného uhla.

- Do 4. sady sme pridali dĺžku osi bunky a maximálny a minimálny priemer rovníka bunky, ktoré sa potenciálne dajú určiť zo snímok z viacerých uhlov.
- 5. sada obsahuje navyše prediktory, ktoré sa dajú dopočítať z kompletnej triangulácie bunky: povrch bunky, objem bunky a priemery, štandardné odchýlky a koeficienty šikmosti vypočítané z dĺžok všetkých hrán triangulácie, uhlov zovretých každými dvoma trojuholníkmi triangulácie a priestorových uhlov pri všetkých uzloch triangulácie. Trianguláciu bunky je z reálneho experimentu získať pomerne náročné, vyžadovalo by to vytvorenie 3D obrazu bunky zo snímaného toku.
- 6. sada obsahuje ešte aj priemery, štandardné odchýlky a koeficienty šikmosti odchýlok a absolútnych odchýlok charakteristík pridaných v 5. sade. Odchýlky sa počítajú od bunky v relaxovanom stave, teda na výpočet potrebujeme mať základnú trianguláciu bunky, čomu pri našom postupe zodpovedá bunka v prvom kroku simulácie.

Uvedené prediktory zodpovedajú aktuálnemu stavu bunky v jednom sledovanom okamihu (v jednom kroku simulácie). Pri experimentoch ale máme možnosť bunku sledovať dlhšie počas niekoľkých krokov. Dá sa očakávať, že presnosť klasifikácie sa zlepší, ak budeme bunku sledovať dlhšie. Aby sme mohli posúdiť tento vplyv, vytvorili sme pre každé $S = 1, 10, 20, 40, 80, 160, 320, 640$ a 1280 model, v ktorom sme každú bunku sledovali S krokov simulácie. (Pozn. pod čiarou: Jeden prechod bunky kanálom zodpovedá približne 100 krokom simulácie.) Keďže sa hodnoty jednotlivých prediktorov počas prechodu bunky kanálom menia, máme pre každý prediktor namiesto jednej hodnoty až S hodnôt. Z týchto hodnôt sme vypočítali priemer a štandardnú odchýlku a tie použili ako prediktory – celkovo tak pre každú

hodnotu uvedenú vo výpise vyššie máme až dva prediktory (s výnimkou prípadu $S = 1$).

Vzhľadom na charakter dát (počet prediktorov, vysoká korelovanosť niektorých prediktorov) za vhodný nástroj pre klasifikáciu považujeme modely založené na rozhodovacích stromoch. Použili sme dve techniky patriace medzi najpopulárnejšie: náhodný les a gradientné rozhodovacie stromy. Pre implementáciu sme použili jazyk Python a metódy `RandomForestClassifier` z knižnice `sklearn` a `XGBClassifier` z knižnice `xgboost`.

Cieľom v tejto časti nebolo maximalizovať presnosť na najvyššiu možnú úroveň, preto sme hyperparametre jednotlivých modelov neoptimalizovali a uspokojili sme sa s predvolenými hodnotami.

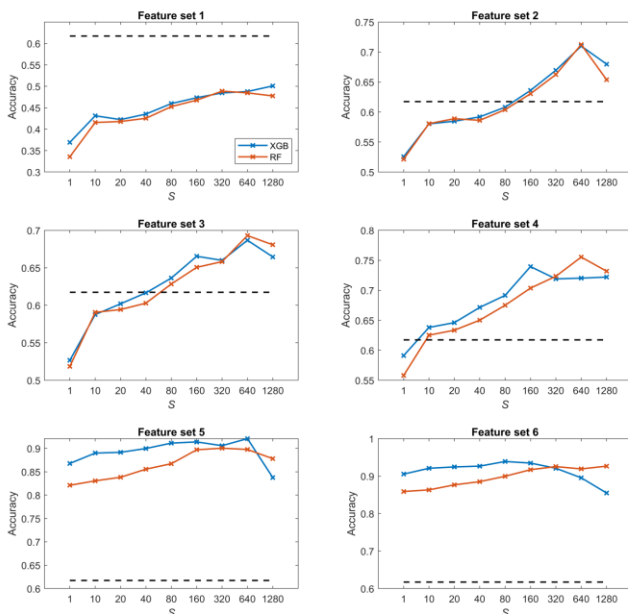
Pre každú zo šiestich sád prediktorov a každú z deviatich hodnôt S sme oboma metódami natrénovali 2 modely – jeden pre klasifikovanie do štyroch tried (štyri stupne tuhosti bunky), druhý pre klasifikovanie do dvoch tried (zdravá/chorá bunka). Pri vyhodnocovaní presnosti klasifikácie do dvoch tried sme okrem druhého modelu použili aj prvý model, pri ktorom sme tri stupne tuhosti zlúčili do výsledku “chorá bunka”.

Zo simulácie sme vytvorili vzorku pre každý krok simulácie, po ktorom nasledovalo aspoň $S - 1$ ďalších krokov (aby sme vedeli vypočítať prediktory pre danú hodnotu S). Odstránili sme však prvých 100 krokov simulácie. Celkový počet vygenerovaných vzoriek je tak rovný $C \times (N - 100 - S + 1)$, kde C je celkový počet buniek v simulácii a N je celkový počet krokov simulácie.

Vzhľadom na použité stromové modely a vynechanie optimalizácie hyperparametrov sme nepotrebovali validačnú časť, takže sme si vystačili s dvoma simuláciami. Tréningové vzorky sme vytvorili zo simulácie s hodnotami $C = 36$ a $N = 2356$, celkový počet tréningových vzoriek je tak v rozsahu od 81216 (pre $S = 1$) po 35172 (pre $S = 1280$). Druhá simulácia, z ktorej

sme vygenerovali testovacie vzorky, má parametre $C = 36$ a $N = 2289$, teda počet vzoriek je v podobnom rozsahu (od 78804 pre $S = 1$ do 32760 pre $S = 1280$).

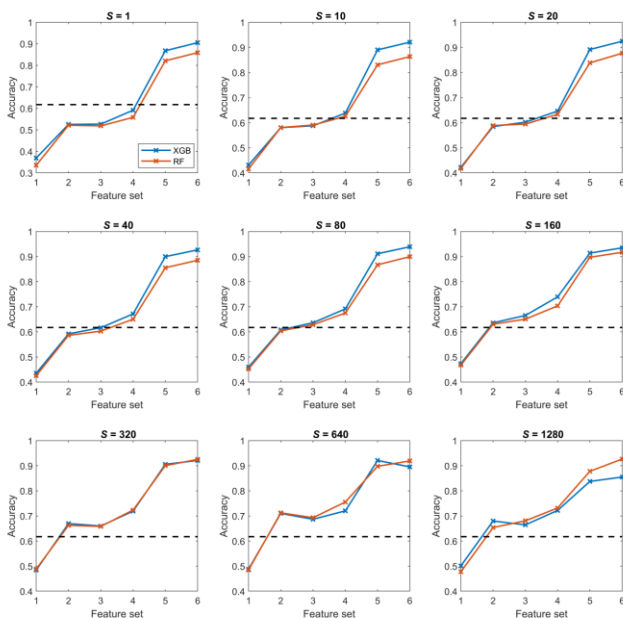
Pozrime sa na presnosť klasifikácie najskôr z hľadiska hodnoty S , teda zaoberajme sa tým, aký vplyv má počet krokov



sledovania jednej bunky. Graf pre klasifikáciu do 4 tried:

Možno usúdiť, že zväčšovanie S má pozitívny vplyv až po úroveň okolo hodnôt $S = 160$ či 320 , od ktorých už sa presnosť klasifikácie prestane kontinuálne zlepšovať. Zdá sa teda, že nemá zmysel bunku sledovať výrazne dlhšie ako počas jedného prechodu kanálom (cca. 100 krokov simulácie).

Presnosť klasifikácie mení podľa toho, ktorú sadu prediktorov použijeme. Najmä pri klasifikácii na 4 triedy možno výrazné zlepšenie presnosti sledovať pri prechode od 4. sady k 5. sade. Naopak, efekt tretieho rozmeru (prechod od 2. sady k 3. sade) sa nezdá byť príliš významný.



Vo všetkých obrázkoch je prerušovanou čiarou znázornená presnosť, ktorú sme dosiahli pri klasifikácii pomocou hlbokých neurónových sietí.

○ **Štatistická analýza elasticity na základe povrchu**

Pre každú z 36 buniek so 4 rôznymi hodnotami elasticity využili sekvenciu 2356 hodnôt jej povrchu a objemu počas pohybu RBC v simulačnom kanály. Každý zápis bol vykonaný po 1000 krokoch simulácie, ktoré zodpovedali posunu RBC

o priemerne 1 až 2 mikrometre v smere x-osi kanála. Z každej dvojice hodnôt SA a V sme následne vypočítali ich pomer SA:V a ďalej pracovali s týmto údajom.

Pre každú RBC zo všetkých 2356 sme z hodnôt pomeru SA:V vybrali len ich štyri základné charakteristiky: maximum, minimum, priemer a rozptyl hodnôt SA:V. Pre každú štatistiku sme tak získali 4 sady po 9 hodnotách pre jednotlivé skupiny rôzne elastických RBC zo simulačného experimentu. V ďalšom sme pokúsili analyzovať, či tieto dátové sady umožňujú rozlíšiť rôznu elasticitu buniek.

Spoločné zobrazenie hodnôt jednotlivých štatistík pre 4x9 RBC naznačuje, že dobré odlišenie normálne elastických RBC od ostatných typov získavame pre hodnoty priemernej SA:V a maximálnej SA:V. Pre hodnoty minima a rozptylu SA:V rozlíšenie nevidíme, a v žiadnej zo štatistík nevidíme rozlíšenie medzi tromi typmi RBC so zníženou elasticitou.

Tento predpoklad sme overili klasickým Kolmogorov-Smirnov testom na hypotézu, že daná sada dát pochádza z toho istého (nešpecifikovaného) rozdelenia pravdepodobnosti. Výsledky sú zhrnuté v nasledovných tabuľkách. Zmienенú hypotézu zamietame pre hodnoty priemerov a hodnoty maxim pri porovnaní buniek typu 0 s bunkami typu 1, 2 alebo 3. Zamietnuť ju môžeme na štandardnej hladine významnosti 5%, pričom ale až na jeden prípad je p-hodnota menej ako 1%.

Aver p-value D-value	type 0	type 1	type 2	type 3
type 0	x	0	0.0007	0.0007
type 1	1	x	0.3517	0.9895
type 2	0.8889	0.444	x	0.7301
type 3	0.8889	0.222	0.333	x

Môžeme teda zhrnúť, že pri využití elementárnych štatistík maxima, minima, priemeru a rozptylu a základného K-S testu

vieme významne odlíšiť priemerné a maximálne hodnoty pomeru SA:V medzi zdravými RBC typu 0 a každou skupinou RBC so zníženou elasticitou. V opačných prípadoch naopak toto rozlíšenie nedosiahneme.

Na druhej strane upozorňujeme na výraznú zhodu získaných záverov s výsledkami spomenutými v časti 1., na ktorých dosiahnutie boli použité výrazne odlišné nástroje a tiež formát a množstvo údajov, aj keď z tej istej východiskovej simulácie.

Pravdepodobnosť potvrdenia výsledku pre simulačné experimenty je možné očakávať aj na základe overenej konzistencie a robustnosti (odkazy?) výpočtového modelu, keď pridávané RBC alebo väčšia dĺžka ich simulovaného toku by nemala priniesť ich výrazne odlišné správanie.

• Záver

V práci sme sa zaoberali tým, ako nám počítačové simulácie pomáhajú pochopiť elasticitu červených krviniek a na predpovedanie ich elasticity sme použili neurónové siete. Simulácie modelovali, ako sa červené krvinky pohybujú v prietoku krvi, čo nám poskytuje dôležité údaje na tréning neurónových sietí. Aj keď simulácie nie sú úplne ako skutočné experimenty, upravili sme výsledky simulácie tak, aby zodpovedali jednoduchým videozáznamom z experimentov.

Zaoberáme sa predpovedaním elasticity červených krviniek pomocou vysokovýkonného CNN-LSTM, pričom berieme do úvahy rôzne dimenzie vstupných údajov. Štúdia zdôrazňuje dôležitosť používania aspoň dvoch zariadení na zaznamenávanie na zachytenie prietoku krvi v lineárne nezávislých rovinách. Architektúra CNN-LSTM s 2D konvolučnými vrstvami je identifikovaná ako najefektívnejšia, dosahuje prijateľnú presnosť, najmä so vstupnými údajmi reprezentujúcimi všetky tri rozmery ohraničujúcich boxov obklopujúcich RBC.

Zameranie ďalej presúvame na výzvy merania rýchlosti deformácie červených krviniek a potenciálu neurónových sietí

kompenzovať stratu presnosti a rozmerov pri spracovaní videozáznamov. Štúdia skúma klasifikáciu elasticity RBC prostredníctvom video analýzy založenej na CNN, pričom využíva architektúry ako ResNet a EfficientNet. Model EfficientNet sa javí ako najlepší klasifikátor a zdôrazňuje sa dôležitosť techník regularizácie a krížovej validácie.

Celkovo práca prispieva k pochopeniu dynamiky červených krviniek a skúma inovatívne diagnostické aplikácie kombináciou počítačových simulácií a videoanalýzy riadenej CNN. Zistenia podčiarkujú potenciál neurónových sietí pri kategorizácii elasticity červených krviniek a zdôrazňujú potrebu starostlivého zváženia rozmerov vstupných údajov a architektúr modelov pre optimálny výkon klasifikácie. Výskum demonštruje efektívnosť tréningu založeného na simulácii, ktorý zaisťuje robustný výkon neurónovej siete aj s obmedzenými súbormi skutočných video údajov.

Na záver naša práca zdôrazňuje potenciál kombinácie počítačových simulácií, videoanalýzy a CNN pre klasifikáciu elasticity RBC. Prebiehajúci výskum a optimalizácia sú nevyhnutné na zvýšenie presnosti, čo v konečnom dôsledku prispieva k zlepšeniu výsledkov zdravotnej starostlivosti.

Zoznam Použitej Literatúry

- [1] Douglas B Kell and Ethersia Pretorius. No effects without causes: the iron dysregulation and dormant microbes hypothesis for chronic, inflammatory diseases. *Biological Reviews*, 93(3):1518–1557, 2018
- [2] Katarína Bachratá, Katarína Buzáková, Michal Chovanec, Hynek Bachratý, Monika Smiešková, and Alžbeta Bohiniková. Classification of red blood cell rigidity from sequence data of blood flow simulations using neural networks. *Symmetry*, 13(6):938, 2021
- [3] Alexander Kovacs, Lukas Exl, Alexander Kornell, Johann Fischbacher, Markus Hovorka, Markus Gusenbauer, Leoni Breth, Harald Oezelt, Dirk Praetorius, Dieter Suess, et al. Magnetostatics and micromagnetics with physics informed neural networks. *Journal of Magnetism and Magnetic Materials*, 548:168951, 2022.
- [4] Alexander Kovacs, Lukas Exl, Alexander Kornell, Johann Fischbacher, Markus Hovorka, Markus Gusenbauer, Leoni Breth, Harald Oezelt, Masao Yano, Noritsugu Sakuma, et al. Conditional physics informed neural networks. *Communications in Nonlinear Science and Numerical Simulation*, 104:106041, 2022.
- [5] Monika Smiešková and Katarína Bachratá. Validation of bulk properties of red blood cells in simulations. In *2019 International Conference on Information and Digital Technologies (IDT)*, pages 417–423. IEEE, 2019.
- [6] Tae-Young Kim and Sung-Bae Cho. Predicting residential energy consumption using cnn-lstm neural networks. *Energy*, 182:72–81, 2019.

[7] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part I 13, pages 818–833. Springer, 2014.

• Zoznam Publikovaných Prác

Molčan, S.; Smiešková, M.; Bachratý, H.; Bachratá, K.: Computational Study of Methods for Determining the Elasticity of Red Blood Cells Using Machine Learning. Symmetry 2022, 14, 1732.

Molčan, S., Smiešková, M., Bachratý, H., Bachratá, K., & Novotný, P.: Classification of Red Blood Cells Using Time-Distributed Convolutional Neural Networks from Simulated Videos. Applied Sciences 2023, 13(13), 7967.

Samuel Molčan ... [et al.] : Models for monitoring progressive multiple sclerosis. Študentská vedecká konferencia FMFI UK, Bratislava 2021, Proceedings of the Student Science Conference 2021 : Zborník príspevkov. 91