Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

Predikčná metóda na určenie klinického dopadu štrukturálnej genomickej variability

Bakalárska práca

2021
Michaela Gažiová

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

Predikčná metóda na určenie klinického dopadu štrukturálnej genomickej variability

Bakalárska práca

Študijný program:	Bioinformatika
Študijný odbor:	Informatika a Biológia
Školiace pracovisko:	Geneton
Školiteľ:	Mgr. Jaroslav Budiš, PhD.

Bratislava, 2021
Michaela Gažiová

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Michaela Gažiová
Študijný program: bioinformatika (Medziodborové štúdium, bakalársky I. st., denná forma)
Študijné odbory: informatika
biológia
Typ záverečnej práce: bakalárska
Jazyk záverečnej práce: slovenský
Sekundárny jazyk: anglický

Názov: Predikčná metóda na určenie klinického dopadu štrukturálnej genomickej variability
Prediction method to determine the clinical impact of structural genomic variability

Anotácia: Moderné sekvenačné technológie umožnili rutinné testovanie na štrukturálnu variabilitu genómu plodu dieťaťa už pred jeho narodením. Zdokonalenie výpočtových metód vedie k detekciám čoraz detailnejších zmien, ktoré sú často jedinečné pre skúmaného jedinca. Určenie, či variácia má klinický dopad na jedinca je teda náročnou úlohou aj pre skúsených klinických genetikov.
Cieľom práce je vyvinúť automatizovaný nástroj na binárnu klasifikáciu detekovanej štrukturálnej variability podľa dopadu na zdravie jedinca. Študentka určí kandidátny zoznam atribútov, ktorými je možné popísať skúmaný variant a určí ich predikčný potenciál. Na trénovanie a testovanie študentka využije publikované záznamy o detekovaných štrukturálnych variáciách z voľne dostupných databáz.

Vedúci: Mgr. Jaroslav Budiš, PhD.
Katedra: FMFI.KI - Katedra informatiky
Vedúci katedry: prof. RNDr. Martin Škoviera, PhD.

Dátum zadania: 29.10.2019

Dátum schválenia: 29.10.2019 doc. Mgr. Bronislava Brejová, PhD.
garant študijného programu

.....
študent

.....
vedúci práce

Poďakovanie:

Ďakujem veľmi pekne môjmu školitelovi Mgr. Jarovi Budišovi, PhD. za nekonečnú trpezlivosť, odborné rady a uvedenie do sveta bioinformatiky. Taktiež aj za nasmerovanie, pripomienky a opravy pri písaní tejto bakalárskej práce.

Ďalej by som rada poďakovala Wernerovi Kramplovi za možnú aj nemožnú podporu, Tomášom, Dii, Ondrejovi&Zuzke, Rasťovi a kolektívu Geneton za odborné rady a príjemnú atmosféru v Genetone.

Ďakujem mojim milým spolužiakom za pomoc pri učení, pozitívnu náladu, ktorú stále šírili a krásne zážitky.

Mojej rodine za podporu pri štúdiu a sestre zvlášť za servis pri písaní patrí veliké ďakujem.

Abstrakt

Štrukturálne varianty (CNV), ktoré sú súčasťou ľudského genómu, môžu spôsobiť závažné vrodené ochorenia. Klasifikácia klinického významu CNV je náročná a zdĺhavá vzhľadom na unikátnosť nálezov a tak množstvo rôznych typov funkčných genomických elementov zasiahnutých daným CNV a ich rozličného dopadu na jedinca. V práci sme sa venovali návrhu automatizovanej predikcie funkčného dopadu CNV. V prvom kroku sme sa zamerali na hľadanie významných anotácií génov prekrytých CNV, ktoré boli získané z rôznych databáz a štúdií pomocou anotačného nástroja AnnotSV. Na trénovanie modelov sme použili zvolené dôležité atribúty a následne predikovali funkčný dopad CNV pomocou viacerých modelov strojového učenia. Vytvorené modely dosiahli úspešnosť na testovacom sete, pozostávajúcom z CNV z ClinVar databázy, 96% na deléciách a 97.3% na duplikáciách. Po aplikovaní stanovených hraníc pre klasifikáciu CNV ako *neklasifikované*, sa táto presnosť zvýšila na 98% pri obidvoch modeloch. V porovnaní s dostupnými nástrojmi, presnosť predikcií výslednej metódy je lepšia alebo znižuje počet CNV klasifikovaných ako *neklasifikované*. Vzhľadom k uvedenému táto metóda by po úpravách mohla urýchliť a pomôcť klinikom pri hodnotení výsledkov genomickej analýzy a následne v procese diagnostiky.

Kľúčové slová: štrukturálny genomický variant, klinický význam, anotácie génov, predikcia patogenicity, algoritmy strojového učenia

Abstract

Copy number variants (CNVs) that are part of the human genome can cause severe diseases. Evaluation of CNV's clinical impact is demanding and difficult due to the variable significance of genomic content which is overlapped by CNV. In this thesis, we aimed to design an automated prediction method for predicting CNV's clinical impact. In the first step, we focused on searching for significant gene annotations overlapped by CNV – gene annotations were obtained from the annotation tool AnnotSV. Some machine learning algorithms were chosen for training on selected attributes. Subsequently, predictions of trained models were compared. Selected models on test sets, consisting of CNVs from the ClinVar database, achieved accuracy of 96% on loss and 97.3% on gains. Moreover, accuracy reached up to 98% on both copy number loss and copy number gain variants when the established thresholds for CNV evaluation as *unclassified* were applied. The accuracy of the model prediction is better or reduces the number of CNVs classified as *uncertain* in comparison to available tools. Thus, after improvements, it would speed up and help clinicians evaluate the results of genomic analysis and diagnostics.

Key words: copy number variant, clinical significance, gene annotations, machine learning algorithms, pathogenicity prediction

Obsah

Zoznam obrázkov

Zoznam tabuliek

Úvod

Rozdiely v štruktúre a zložení DNA sú zdrojom ľudskej genetickej variability. Takýmito rozdielmi sú aj *štrukturálne varianty typu CNV (copy number variants)*, ktoré zahŕňajú zduplikovanie alebo odstránenie segmentu chromozómu. Tie možu mať výrazný dopad na molekulárny ako aj následný fenotypový prejav. Deléciou génov či regulačných oblastí môže nastať zníženie až absencia génového produktu, duplikáciou zase nadbytok proteínu v organizme. Zasiahnutie kritických regiónov môže vyústiť k vzniku vrodených ochorení. Pre klinickú diagnostiku je preto dôležitá včasná detekcia a určenie ich klinického významu. Táto klasifikácia je ale problematická kvôli rôznym, často unikátnym kombináciám genomických elementov, ktoré sú prekryté daným CNV. Nakoľko dostupné klasifikačné nástroje (klasifikujúce napr. na základe expertných skórovacích schém) veľké množstvo CNV klasifikujú ako *nejasný* význam a je potreba klinického experta, ktorý by mal potrebné vedomosti na vyhodnotenie častí, ktoré nemôžu byť zautoamtizované, a tak ukončiť klasifikáciu dopadu, navrhli sme nástroj na automatizovanú predikciu klinického dopadu CNV, s cieľom spresniť a zjednodušiť náročnú úlohu klinickým diagnostikom pri hodnotení výsledkov genomickej analýzy a klasifikácii. Implementovaný prediktor vyžaduje na vstupe iba pozíciu CNV (pozícia od-do na konkrétnom chromozóme) a jeho typ (delécia alebo duplikácia). Výstupom je pravdepodobnosť patogenicity a klasifikácia dopadu na organizmus – neškodný (benígny) alebo klinicky závažný, spôsobujúci vrodené genetické ochorenie (patogénny).

V prvej kapitole sa zoznámime s potrebnými biologickými pojmami, ako aj možnými dopadmi prítomnosti CNV (časť 1.2) a zadefinujeme ciele práce (časť 1.4). V časti 1.3 si predstavíme dostupné nástroje klasifikujúce klinický dopad CNV a popíšeme metodiku ACMG klasifikačných kritérií. Ďalej sa oboznámime s dostupnými dátovými zdrojmi a ich atribútmi. V kapitole *Metódy* si pripravíme dátové sady zozbieraných verejne dostupných delécií a duplikácií so známym klinickým dopadom (v časti 2.1) a porovnáme hodnoty patogénnych a benígnych atribútov pomocou štatistických metód. Ďalej v časti 2.2 si z informatického hľadiska zadefinujeme klasifikačný problém a oboznámime sa s fungovaním niektorých algoritmov strojového učenia. Porovnáme predikcie modelov na pripravených dátach a zadefinujeme štatistické metódy na vyhodnocovanie správností testov na základe ktorých budeme tieto predikcie porovnávať. Nakoniec v časti 2.2.5

predstavíme postup na identifikáciu hraníc predikcie patogenicity pre klasifikáciu CNV ako *neklasifikované*. V poslednej kapitole zhrniem výsledný výber a použitie atribútov v modeloch pre delécie a duplikácie (časť 3.1). Porovnáme presnosti modelov a zvolíme vhodný model (časť 3.2). Ďalej určíme výsledné hranice predikcie patogenicity CNV pre klasifikáciu ako *neklasifikované* a dopad na predikcie v časti 3.3. Napokon v časti 3.5 porovnáme výsledky predikcií modelu s dostupnými klasifikačnými nástrojmi.

Kapitola 1

Úvod do problematiky

1.1 Biologický základ

V tejto podkapitole si zadefinujeme potrebné pojmy, ktoré v texte budeme používať z hľadiska biológie.
Nositeľom genetickej informácie každého jedinca je *DNA (deoxyribonukleová kyselina)*. Toto vlákno (reťazec) je tvorené nukleotidmi, ktorých súčasťou sú dusíkaté bázy *A, C, G, T*.

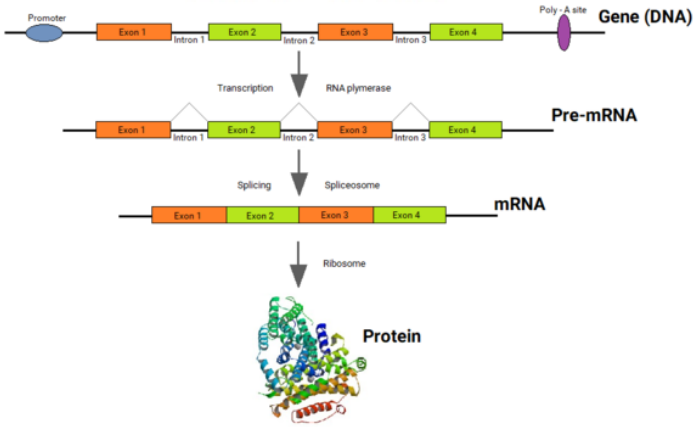
Pri replikácii DNA (napr. pri delení bunky) sa lineárna DNA za pomoci rôznych proteínov kondenzuje – zbaľuje do kompaktnej štruktúry, do štruktúry *chromozómov*. Pohlavné bunky majú 23 chromozómov. Somatické bunky 23 párov chromozómov, pričom obsahuje po dvoch kópiách z každého chromozómu. Jedna kópia zdedená od matky a druhá od otca. Tieto páry nazývame homologickými chromozómami – homológmi alebo autozómami. Heterochromozómami sú pohlavné chromozómy – gonozómy X a Y. [1]

Pojmom *genóm* budeme uvažovať kompletný súbor informácií v jadrovej DNA na autozómoch a gonozómoch daného organizmu.

Úsekom kódujúcim informáciu pre tvorbu produktu (proteínov) sú *gény*. Sekvencia génu obsahuje *exómy* – kódujúca časť a *intróny*, ktoré sú nekódujúce sekvencie medzi exómami jedného génu. Splicing (vystrihnutie intrónov) môže spôsobiť aj vynechanie niektorých exómov. [1] Ďalšími nekódujúcimi zložkami DNA sú *regulačné oblasti*:

- promotóry – obsahujú časti (napr. TATA box, CAT box), na ktoré sa viažu transkripčné faktory - tie sa podieľajú na inicializácii transkripcie, keďže naň nasadá RNA polymeráza a následne začína transkripcia.

- enhancery a silencery – ovplyvňujú transcripciu génov. Enhancery ju zvyšujú pri nasadaní aktivačných faktorov, a tým že viažu RNA polymerázu, spúšťajú transkripciu. Silencery zabraňujú transkripcii, keďže viažu inhibičné faktory.



Obr. 1.1: Expresia génu a proces tvorby proteínu z génu. Zdroj: wikipedia.org

DNA sekvencia obsahuje regulačnú oblasť (promotór), gén je zložený z intrónov a exómov. Exómy sú kódujúcou zložkou, ktoré nesú informáciu pre vznik proteínu.

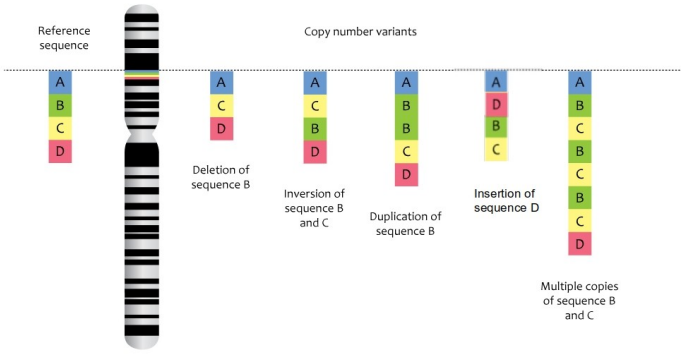
- operátor – táto sekvencia sa nachádza medzi promotórom a sekvenciou génov. Pri nasadnutí aktivovaného represoru na operátor je znemožnená transkripcia, nakoľko represor bráni pokračovaniu transkripcie po nasadnutí RNA polymerázy na promotór.
- terminátor – ukončuje transkripciu génu tým, že po transkripcii tohto úseku (poly adenylový koniec) sa RNA polymeráza odpája z vlákna DNA.

Expresia génu sa podieľa na procese vzniku produktu génov a tento priebeh je znázornený na obr. 1.1. Regulačné oblasti ovplyvňujú expresiu génu, keďže zabraňujú alebo podporujú iniciáciu či pokračovanie transkripcie.

Organizmy sa medzi sebou líšia, pretože sa líši ich postupnosť nukleotidov. Odlišujú sa najmä preto, lebo DNA nesie informáciu o génoch, ale aj o regulačných faktoroch génovej expresie.

Rozdiely v štruktúre chromozómov sú zdrojom ľudskej genetickej variability. Takýmito rozdielmi sú aj *štrukturálne varianty (SV)*, ktoré zahŕňajú rôzne genomické zmeny ako sú delécie, duplikácie, inzercie, inverzie alebo translokácie. Typy variantov sú znázornené na obr. 1.2. Dĺžka takýchto úsekov je väčšia ako 50 bp, avšak táto spodná hranica je variabilná – považuje sa aj 1000 bp [42].

Zo štrukturálnych variantov sa budeme zaoberať deléciami (copy number *loss*) a duplikáciami (copy number *gain*) dlhšími ako 1000bp, ktoré budeme súhrne označovať ako CNV varianty (copy number variants). Pri CNV variantoch dochádza ku zmenám počtu kópií určitého úseku v genóme, pričom sa môže tento počet ľubovoľne znásobiť. Väčšinou však sú to delécie, pričom do úvahy berieme vymazanie úseku aspoň na jednej kópii chromozómu, alebo duplikácie úseku aspoň na jednej kópii chromozómu. Vyskytujú sa ale aj prípady viacnásobnej zmeny, ako sú triplikácie, tetraplikácie, či



Obr. 1.2: Typy variantov. Zdroj: Upravený Obrázok z upload.wikimedia.org/wikipedia/commons/2/2b/Copy_number_variants.png

väčšie násobky úsekov. [42] Pre duplikácie uvažujeme až do päťnásobného počtu kópií úseku.

1.2 Popis problému

Naproti jednonukleotidovým polymorfizmom (SNP), ktoré sú častejšie, majú tieto štrukturálne varianty väčší funkčný potenciál v dôsledku svojej väčšej veľkosti a je pravdepodobnejšie, že spôsobia zmenu v expresii génov. [32]

CNV môžu prispieť k rozvoju ochorení:

- priamo narušením samotného génu, čím nastane zmena a vytvorí sa nová funkcia produktu génu. Pri delécii úseku na ktorom sa nachádza gén, bude nám jeho produkt chýbať. V prípade, že je detekované CNV duplikáciou, tak génu a následne nasyntetizovaného proteínu bude v organizme veľa. Ak sa s týmto nadbytkom nebude vedieť vysporiadať, tak je pravdepodobné, že nastane problém.

- nepriamo zasiahnutím regulačnej oblasti, čím sa naruší expresia prislúchajúceho génu – teda zmena množstva vyprodukovaného proteínu, či už viac, menej alebo vôbec sa nebude produkovať, čo spôsobí absenciu konkrétneho produktu.

So zmenami spôsobených CNV súvisia aj tieto dva pojmy:

Haploinsuficiencia je situáciou, ktorá nastane, keď je jedna kópia génu vymazaná a zostávajúca funkčná kópia génu nie je schopná vytvoriť dostatočné množstvo génového produktu [25].

Pri *triplosensitivite* je problémom nadmerná tvorba génového produktu, čo spôsobuje problém v organizme [13].

Tieto molekulárne následky CNV môžu vyústiť do rôznych ochorení, syndrómov, malformácií, či možného vzniku rakoviny. Z mikrodelečných syndrómov sú to napr.

1p36, DiGeorgov, Cri du Chat, Prader-Willi, Wolf-Hirschhornov syndróm. [40] Duplikácie v dôsledku nadmernej expresie génov PLAGL1 a HYMAI môžu spôsobiť aj ochorenie diabetes mellitus [26]. Možné ochorenia sú spojené s vývinovými, kardiovaskulárnymi, neurodegeneratívnymi, či autoimunitnými poruchami. [40]

V prípade, že tieto CNV môžu spôsobiť ochorenia a tým výrazne znížiť kvalitu a dĺžku života, hovoríme, že môžu byť *patogénnymi* CNV. V opačnom prípade, ak sa nepodieľajú na vzniku ochorení, jedná sa o *benígne* CNV. Dopady benígneho CNV sú prevažne neutrálne, avšak v niektorých prípadoch môžu zlepšiť schopnosť prežitia v nepriaznivom prostredí [40].

Motivácia Pre spomenuté možné následky prítomnosti CNV je dôležité vopred identifikovať možný dopad nájdeného CNV. Aj keď existujú také záznamy o klinickom význame niektorých CNV z rôznych štúdií, veľa z nich je zaradených do kategórie *nejasného významu* [36]. Kľúčovým problémom je identifikovanie kauzálneho variantu zapríčiňujúceho chorobný prejav. Taktiež, pri prenatálnom testovaní alebo skríningu novonarodených je dôležité určiť, či bude CNV potenciálne problémové. Naliehavou výzvou je preto ich presná klasifikácia, a tak včasné zahájenie a správne zacielenie liečby.

1.3 Súčasný stav riešenej problematiky

V tejto časti bude prezentovaný prehľad a popis podobných existujúcich metód na predikciu patogenity CNV, klasifikáciu CNV a nástrojov anotujúcich CNV.

Väčšina dostupných nástrojov sú klasifikátory, ktoré klasifikujú klinický dopad CNV na základe expertmi navrhnutých *ACMG kritérií*. Pri nich sa jedná o jednoduchú skórovaciu schému, ktorá má 5 základných kategórií. Prvá hovorí o prítomnosti proteín kódujúcich génov či iných známych dôležitých elementov, ktoré sú prekryté daným CNV. Druhá sa zameriava na prekrývanie haploinsuficientných génov a regiónov s CNV alebo, či sa CNV prekrýva so známymi benígnymi CNV. Tretia udeľuje skóre stanovené podľa počtu proteín kódujúcich génov prekrytých alebo pokrytých CNV. O podrobných hodnoteniach genomického obsahu CNV z publikovanej literatúry, štúdií a dostupných databáz hovorí štvrtá kategória. Posledná – piata kategória priraďuje skóre na základe vyhodnotení rodinnej anamnézy pacienta, či je CNV zdedené alebo vzniknuté de novo. Skóre podľa vybraných možností z jednotlivých častí kategórií sa zrátajú.

Podintervaly výsledného skóre následne klasifikujú klinický prejav CNV na *benígne* (≤ -0.99), *pravdepodobne benígne* (od -0.98 do -0.90), *nejasný význam (VUS)* (od -0.90 do 0.90), *pravdepodobne patogénne* (od 0.90 do 0.98) a *patogénne* (≥ 0.99). [44]

Avšak, nevýhodou ACMG kritérií je potreba klinického experta, ktorý by veľké množstvo času strávil pri prehľadávaní rôznych zdrojov a publikácií, aby ohodnotil jednotlivé kategórie a určil výsledné skóre pre vyhodnotenie ACMG kritériami. Problematické je aj pridelenie skóre pre danú časť, nakoľko pre viacero možností je možné skóre upraviť podľa subjektívneho úsudku klinika. Viacero klinikov teda na základe týchto kritérií môže dospieť ku konfliktným klasifikáciám CNV nálezov.

Jednotlivé nástroje klasifikujúce podľa ACMG kritérií sa líšia v použití dátových zdrojov, ohodnotení, a teda aj následnom vyhodnotení (klasifikovaní) klinického prejavu.

ClassifyCNV

je nástrojom, ktorý klasifikuje automatizovaným výberom možností jednotlivých kategórií ACMG kritérií. Informácie o génoch a promotóroch získava z *refGene* databázy, o triplosensitívnych génoch z *ClinGen*–u, o haploinsuficientných génoch z *ClinGen*–u a *DECIPHER*–u, o enhanceroch (regulačných oblastiach) z *VISTA, FANTOMS, Ensemble* databáz.

Problémom pri automatizovanom vyhodnotení ACMG kritérií je, že nie je možné z voľne dostupných databáz ohodnotiť všetky kategórie ako sú napr. rodinná anamnéza, či bližšie informácie o podobných prípadoch z publikovanej literatúry. Výstupom je súbor s hodnotami pre výpočet overerených postupov ACMG kritérií a klasifikáciou pravdepodobného klinického prejavu pre zadané CNV. Tento nástroj je dostupný na platforme GitHub: [18]. [19]

AnnotSV

AnnotSV je taktiež nástrojom, ktorý klasifikuje podľa ACMG kritérií, pričom predchádzajúca verzia klasifikovala SV podľa vlastných stanovených kritérií. Informácie o génoch, transkriptoch a kódujúcich oblastí získava z databázy *RefSeq*, populačné frekvencie génov z *1000genomes project* a *gnomAD*–u, informácie o triplosenzitivite a haploinsuficiencii génov z *ClinGen*–u, či názvy chorôb z *OMIM*–u a *DDD*, ktoré sú pridružené pri výskyte prekrývajúceho CNV s daným CNV.

Výstupom AnnotSV je súbor obsahujúci anotácie CNV ako celku a aj po jednotlivých génoch, ktoré sú zasiahnuté daným variantom. Pre gény a celkové CNV poskytuje minimálne 96 anotácií zo spomenutých databáz a nakoniec skóre podľa ACMG kritérií a prislúchajúcu klasifikáciu klinického prejavu.

Možnosť použitia *AnnotSV* nástroja je cez webové rozhranie (na ich stránke [17]) ako aj v samostatne inštalovateľnej aplikácii. [16]

AluScanCNV2

Tento prediktor je postavený na strojovom učení a predikuje náchylnosť CNV byť rakovinotvorným. Okrem spomínanej predikcie umožňuje aj detekciu CNV (*cnv calling*) z *NGS* (Next generation sequencing) výstupov. Tento balík implementovaný v jazyku *R* je dostupný na platforme GitHub: [23]. [24]

1.4 Cieľ práce

Vzhľadom k vyššie spomenutým možným následkom prítomnosti CNV, dôležitosti klasifikovania klinického významu CNV, a nakoľko existujúce nástroje nie sú dostatočne presné alebo veľkú časť hodnotia ako *nejasný význam*, a tiež nedostatočným softvérovým vybavením určeným na predikciu patogenicity CNV a s tým spojenou potrebou klinického experta, ktorý by mal potrebné vedomosti a mnoho času na vyhodnotenie, rozhodli sme sa vytvoriť predikčnú metódu na určenie klinického dopadu CNV určenú pre klinických diagnostikov.

Na vstupe dostaneme úsek CNV s pozíciou prvej a poslednej bázy CNV nálezu, číslo chromozómu a typ CNV, pričom uvažujeme iba jeho deléciu alebo duplikáciu. Výstupom bude pravdepodobnosť s akou je dané CNV patogénne a klasifikácia podľa tejto predikcie na benígne, neklasifikované a patogénne.

Anotácie CNV a génov budeme čerpať z výstupu nástroja AnnotSV [16]. Oboznámime sa aj s dostupnými dátovými zdrojmi a anotáciami genomického regiónu, ktoré ponúkajú. Nakoniec vyberieme vhodné anotácie, ktoré korelujú so závažnosťou CNV nálezu, ktoré využijeme na trénovanie modelu. V práci popíšeme algoritmy strojového učenia, ktoré využijeme na natrénovanie modelov na predikciu patogenicity CNV. Určíme hranice predikcie vybraného modelu na klasifikovanie CNV ako *neklasifikovaných*. V závere práce vyhodnotíme presnosti modelu na pripravených testovacích dátach a výsledky porovnáme s klasifikačnými nástrojmi ClassifyCNV a AnnotSV.

1.5 Databázy variantov a génov

Pre vyhodnotenie závažnosti CNV sa využívajú viaceré genomické databázy, ktoré využívajú aj automatizované nástroje. Anotácie z týchto zdrojov je možné automatizovane získavať pomocou nástroja AnnotSV.

ExAC

Databáza ExAC (Exome Aggregation Consortium) sa zameriavala na poskytnutie informácií o bodových variantoch proteín kódujúcich génov, ktoré sú v populáciách v nízkej frekvencii. Boli zozbierané z exómových údajov sekvenovania 60 706 jedincov

z celosvetovej populácie. Táto štúdia analyzovala citlivosť génov pri prítomnosti variantov – synonymických, missense, delécií, duplikácií. Identifikovala gény, ktoré sú pri ich mutácii náchylné na stratu funkcie a tieto ešte neboli zaradené do OMIM-u, či ClinVar-u ako pozorované varianty génu. Ďalej identifikovali gény, ktoré pri zvýšení alebo znížení množstva daného génu (pri zmenenom počte kópií) spôsobia problém. Takisto určili aj alelové frekvencie variantov proteín kódujúcich génov pre jednotlivé populácie. [33]

V súčasnosti časť z týchto zozbieraných informácií je súčasťou gnomAD-u.

ClinVar

ClinVar je voľne prístupnou databázou ľudských variantov. Zahŕňa poznatky o vzťahoch medzi variantami, ale aj interpretáciu fenotypov (klinického významu) pozorovaných na zdravotnom stave pacienta. Klinický význam variantov je hodnotený podobne ako pri ACMG kritériách ako benígne, pravdepodobne benígne, nejasný prejav, pravdepodobne patogénne a patogénne. K parametrom CNV patrí aj skóre reprezentujúce mieru, ktorou je klinický význam podporený (napr. zohľadnené konfliktné záznamy o klinickom význame CNV), ako aj počet prispievateľov. [31]

DGV

Databáza genomických variantov (DGV) poskytuje štrukturálne varianty, ktoré sa nachádzajú v genómoch jedincov zozbieraných z viacerých globálnych populácií. Obsahuje viac ako 2 500 000 záznamov SV identifikovaných v 22 300 genómoch, ktoré zozbierali z 55 publikovaných štúdií. Ďalšie údaje získali z archívnych databáz SV *dbVar* (NCBI) a *DGVa* (EBI), pričom správnosť týchto údajov najprv ohodnotili. [34]

DDD

Štúdia The Deciphering Developmental Disorders (DDD) zozbierala dáta sekvencií exómov u detských pacientov, u ktorých sa vyskytli závažné vývinové poruchy, ktoré lekári nevedeli diagnostikovať, a exómy ich rodičov z Veľkej Británie a Írska. Zisťovali molekulárnu podstatu, ktoré vysvetľujú dané poruchy u detí, čiže funkciu génov podieľajúcich sa na príčine, vývoji a úlohe daného fenotypu. Tieto výsledky sú dostupné v databáze DECIPHER [14] a v European Genome-phenome Archive. [15]

ClinGen

Clinical Genome Resource (ClinGen) konzorcium hodnotí, či zmeny v génoch môžu spôsobiť ochorenie a určujú, aké varianty génov môžu mať závažný klinický dopad, či zvýšenie alebo zníženie množstva daného génu alebo genomického regiónu môže

vyústiť do vzniku ochorenia. Poskytuje tiež klinickú interpretáciu variantov spojených s rakovinou a ponúka anotácie údajov z publikovanej literatúry. [43]

OMIM

Databáza OMIM (Online Mendelian Inheritance in Man) [20] ponúka súhrnné a podrobné informácie o génoch, genetických fenotypoch a ochoreniach, a predovšetkým sa zameriava na vzťah fenotypu a genotypu. Navyše existuje k nej aj prístupný vyhľadávač na web stránke `omim.org` (*An Online Catalog of Human Genes and Genetic Disorders*), ktorý ku génom obsahuje: MIM identifikačné číslo, alternatívne názvy, označenie génu, pozíciu (cytogenetickú lokáciu aj genomické koordináty), pridružené ochorenie, či syndróm, funkciu génu, molekulárny význam (čo gén kóduje), pozorované varianty génu a ich molekulárny a fenotypový popis, a tým pádom aj množstvo zdrojov na publikovanú literatúru popisujúce súvisiace prípady. Taktiež ku ochoreniam a syndrómom sú pridané alternatívne názvy, gény zapojené do problému vzniku ochorenia, popis ochorenia, prejav (aké mutácie proteínov sa vytvoria, kde nastane problém, symptómy), popis vykonaných laboratórnych testov, dedičnosť ochorenia, možné varianty génov znižujúce riziko ochorenia a asociácie s inými génmi. Tento vzťah genotyp – fenotyp predstavuje vytvorenú *Morbid Map*u. Naviac, OMIM ponúka aj externé zdroje na iné databázy, ktoré zahŕňajú skúmaný gén, proteín či pridružené ochorenie. [2]

GnomAD

GnomAD (Genome Aggregation Database) je agregáciou 125 748 exómov a 15 708 genómov z viacerých výskumných štúdií. Súčasťou je 241 000 000 variantov dĺžky do 50 bp a 335 470 SV. Týmto rôznym druhom variantov priradili aj potenciálny funkčný vplyv. Nakoľko tieto zdroje sú získané z globálnej populácie (54% ne-Európska populácia), obsahuje aj alelové frekvencie variantov pre jednotlivé populácie. Obsahuje aj gény u ktorých zistili, že aj keď obe kópie génu sú neaktívne alebo vymazané, strata funkcie génu nespôsobuje problém. Taktiež tu nájdeme pravdepodobnosť straty funkcie génu následkom porušenia regulačnej oblasti toho génu.[28]

Kapitola 2

Metódy

Na vstupe máme informácie o vyhodnocovanom CNV: číslo chromozómu (1, 2, 3... 22, príp. X/Y), pozície prvej a poslednej bázy CNV na tomto chromozóme, typ CNV, či je duplikáciou alebo deléciou. Dané CNV najprv oanotujeme potrebnými atribútmi z vyššie uvedených štúdií z predchádzajúcej kapitoly 1.5 pomocou nástroja AnnotSV popísaného v časti 1.3. Následne spustíme na týchto dátach natrénovaný predikčný model klasifikátora, ktorý vyhodnotí, s akou pravdepodobnosťou bude dané CNV patogénne.

2.1 Príprava dát

Dáta, s ktorými sme pracovali sú voľne dostupné a stiahnuteľné z ClinVar databázy [31] (stiahnuteľné napr. tu [12]). Celkový počet variantov bol 46 534. Z nich sme si vybrali tie delécie a duplikácie, ktoré boli v ClinVar databáze hodnotené ako patogénne alebo benígne a ich dĺžka bola väčšia ako 1000bp. Tento počet sa nám zredukoval na 28 892 CNV (16 840 delécií, 12 052 duplikácií). Ďalej na anotovanie týchto variantov sme použili nástroj AnnotSV [16] (popísaný v časti 1.3). Vstupnými dátami pre AnnotSV boli dáta z ClinVar databázy, ktoré sme si vyfiltrovali ako je vyššie spomenuté. Vstupom bol súbor v bed formáte, ktorý obsahoval číslo chromozómu, počiatočnú a koncovú pozíciu CNV, typ o aké CNV sa jedná (či je duplikáciou alebo deléciou), identifikátor RCV (identifikátor určujúci CNV v ClinVar databáze) a iné parametre pre CNV, ktoré boli súčasťou CNV v databáze ClinVar. Výstupom AnnotSV bol súbor obsahujúci anotáciu a ohodnotenie štrukturálneho variantu ako celku a aj po jednotlivých génoch, ktoré sú zasiahnuté daným variantom.

Následne CNV s týmito dátami boli rozdelené do troch dátových setov. V pomere trénovací 70%, validačný 15% a testovací set 15%.

Pre prácu s tabuľkovými dátami sme použili knižnicu Pandas [48].

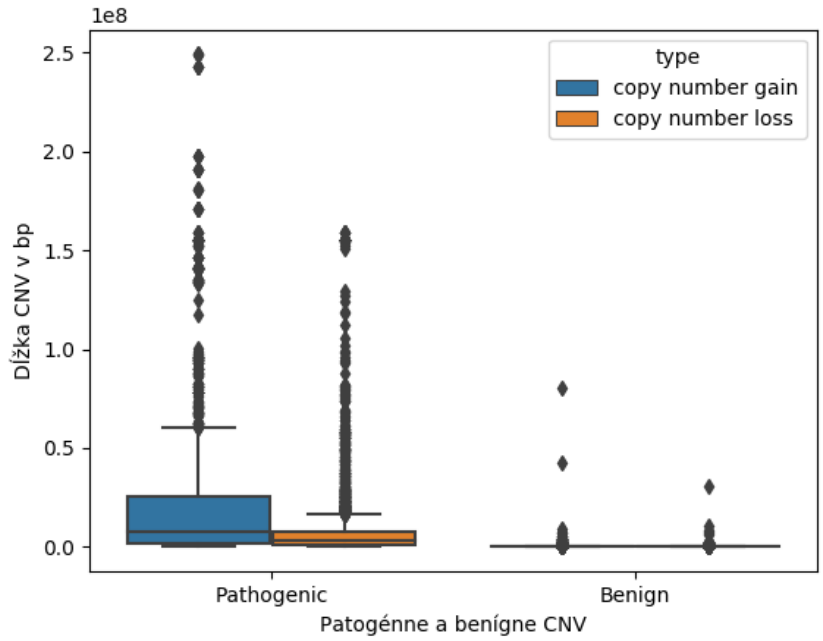
2.1.1 Hľadanie a ohodnotenie významných atribútov

Najprv sme si trénovacie dáta rozdelili na patogénne (4180) a benígne (7608) delécie a podobne duplikácie na benígne (6417) a patogénne (2019). Z výstupu AnnotSV sme vybrali anotácie CNV a génov, ktoré obsahovali číselné hodnoty. Vybrali sme len také hodnoty, keďže ich budeme neskôr porovnávať (takých atribútov bolo 58). Spozorovali sme tiež, že aj atribúty, ktoré obsahujú text sú z molekulárneho hľadiska významné a nesú cennú informáciu. Na základe toho sme sa rozhodli, že vytvoríme nové atribúty CNV vychádzajúc z počtu génov, ktoré CNV prekrývajú a nesú konkrétnu informáciu (proteín kódujúce gény, morbídne gény, triplosenzitívne a haploinsuficientné gény). Taktiež aj podľa pozorovaného zastúpenia alelových frekvencií génov sme vytvorili CNV atribút výberom najnižšej z frekvencií génov, či maximálne Z-skóre nefunkčnosti génu, zo všetkých génov prekrytých CNV.

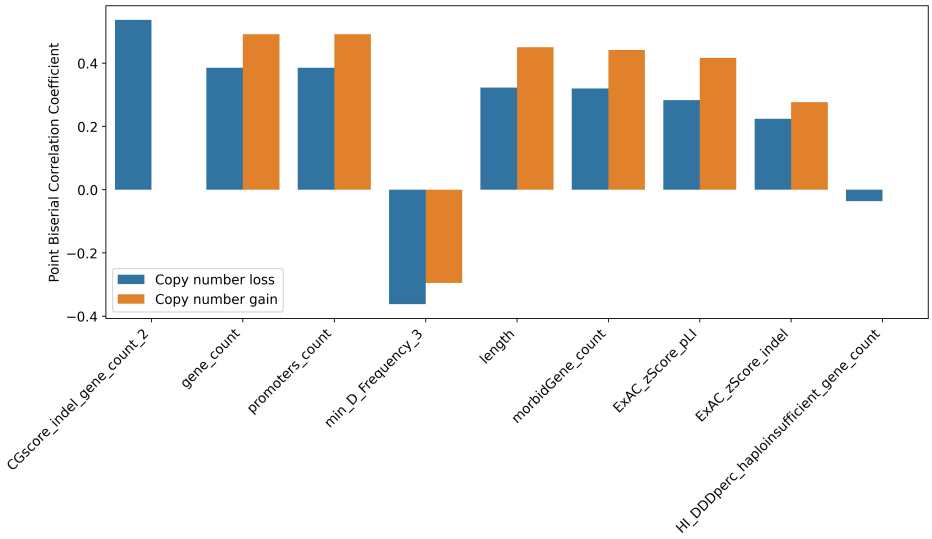
Aby sme zistili, ktoré sú významné atribúty, ktoré korelujú s benígnym/patogénnym CNV otestovali sme rozdiely v atribútoch pomocou štatistického testu Mann-Whitney U Testu [35]. Túto štatistickú metódu sme sa rozhodli použiť, keďže porovnávame dva číselné vektory atribútov pre patogénne a benígne CNV. Tieto vektory nie sú normálne distribuované (každý atribút dosiahol p-hodnotu menšiu ako 0.001 podľa D'Agostino testu normality – použili sme implementovaný v knižnici scipy.stats.normaltest [27]), vybrali sme preto tento neparametrický test. Test sa využíva pri porovnávaní stredných hodnôt dvoch nezávislých vzoriek a určuje, či je rozdiel stredných hodnôt dvoch skupín štatisticky významný alebo náhodný [45]. My sme testovali rozdiely medzi dvomi vektormi jedného atribútu reprezentujúce rôzne typy nálezov – skupina benígnych a skupina patogénnych nálezov. Použitý test je implementovaný v knižnici scipy.stats.mannwhitneyu [27]. Z výsledkov týchto porovnávaní atribútov sme vybrali také, ktorých p-hodnoty vyšli hlboko pod konzervatívnu hranicu štatistickej významnosti (0.001), kedy je nízka šanca, že rozdiely v hodnotách atribútoch patogénnych a benígnych CNV nie sú náhodné. Počet atribútov sa zredukoval z 58 na 42 atribútov. Vizualizácie rozdielností pre tie atribúty, ktorých p-hodnota podľa Mann-Whitney testu [35] bola menšia ako 0.001, sú na grafoch v elektronickej prílohe. Na ukážku porovnanie dĺžok na obr. 2.1.

Atribúty sme takisto porovnali pomocou Point-Biserial korelačného koeficientu, ktorý predstavuje mieru korelácie medzi vektorom hodnôt a vektorom obsahujúcim dve premenné – v našom prípade je to klinický význam CNV (patogénne alebo benígne). Interpretácia výstupu je podobná ako pri Pearsonovom korelačnom koeficiente. [7]. Použitý test je implementovaný v knižnici scipy.stats.pointbiserialr [27]. Vizualizácia na obr. 2.2.

Jednotlivé atribúty sme tiež porovnávali medzi sebou pomocou Pearsonovho korelačného koeficientu. Ten označuje mieru lineárnej korelácie medzi dvomi vektormi a



Obr. 2.1: Porovnanie dĺžok patogénnych a benígnych CNV
Porovnanie patogénnych a benígnych CNV podľa dĺžky.



Obr. 2.2: Porovnanie CNV atribútov pomocou Point-Biserial korelačného koeficientu pre delécie (modré) a duplikácie (oranžové)

nadobúda hodnoty z intervalu < -1, 1>. Lineárny vzťah medzi skúmanými veličinami nie je, ak je korelačný koeficient rovný 0. Pri rovnom jednej, veličiny sú závislé, a teda pri zvýšení jednej veličiny sa zvýši aj druhá. Naopak, ak je korelačný koeficient rovný -1, tak zvýšenie jednej spôsobí zníženie druhej veličiny. Vizualizácie sú na obrázkoch v elektronickej prílohe.

Na vykresľovanie grafov sme využili knižnicu Seaborn [47]. Všetky korelácie patogénnych a benígnych CNV sú k nahliadnutiu na priložených grafoch v elektronickej prílohe a v súbore correlations_info_train_loss_gain.tsv, ktorý obsahuje celkové hodnotenie atribútov pre delécie aj duplikácie.

2.2 Modelovanie

V tejto podkapitole sa budeme venovať rozdielom rôznych klasifikačných metód a výberom klasifikátora.

Pozrieme sa ako niektoré fungujú a nakoniec ich funkčnosť porovnáme na pripravených dátach (časť 2.2.3).

Modelom budeme označovať reprezentáciu, ktorá je naučená (natrénovaná) zo zadaných dát algoritmom uskutočňujúcim tento proces učenia (trénovania) modelu.

$$Model = Algoritmus(Dáta)$$

Predikívne modelovanie zahŕňa strojové učenie, rozpoznávanie vzorov, či metód a techník na vyhľadanie súvislostí v dátach [30]. Zaraďuje sa teda k najbežnejším typom strojového učenia.

J. Brownlee vo svojej knihe [9] definuje, že "algoritmy strojového učenia odhadujú mapovaciu funkciu f výstupu Y daných vstupných premenných X.", môžeme zapísať ako $Y = f(X)$. To tiež predstavuje, že po aplikovaní konkrétnych operácií na vstupné hodnoty získame výstup v podobe predikcie. Atribútmi vstupných hodnôt budeme nazývať aj tzv. *features*. A teda pri prediktívnom modelovaní to môžeme zadefinovať ako

$$Predikcia = Model(InputFeatures)$$

Typom strojového učenia, ktorým sa budeme venovať sa nazýva *supervised machine learning (učenie s učiteľom)* – pri ňom sa model učí (trénuje) na základe zozbieraných a správne určených dát a následne predikuje daný jav, s ktorým sa ešte nestretol [30]. Ku základným typom tohto problému patrí klasifikácia a regresia. Hlavným rozdielom týchto dvoch problémov je, že pri klasifikácii sa zo vstupných údajov predikuje kategória (kategorický label), pričom pri regresii je predikovaná spojitá hodnota - predikuje kvantitu [22].

2.2.1 Klasifikácia, Regresia

Pri *klasifikácii* chceme pre vstupné údaje zistiť, ku ktorej z k kategórií patrí. Pri tomto probléme algoritmus vytvorí funkciu $f : R^n \to \{1, \ldots, k\}$ a následne môže byť použitá na vyhodnotenie nového vstupu. Obdobným výstupom klasifikačného problému môže byť distribúcia pravdepodobností pre dané kategórie. [4]

Binárnou klasifikáciou nazývame prípad, keď sa klasifikuje do dvoch kategórií. Ak ide o problém s viac ako dvomi kategóriami (labelmi) hovoríme o *multi-class klasifikácii*. [10]

Pri *regresii* sa zo vstupu predikuje numerická hodnota. Pri tomto probléme algoritmus vytvorí funkciu $f : R^n \to R$ [4]. Príkladom regresie môže byť predikcia množstiev, či veľkostí.

Rozdielom s klasifikáciou je vo výstupe, keďže u regresii sú výstupom spojité hodnoty a u klasifikácii sú to diskrétne hodnoty (ako/alebo kategórie). [10]

Niektoré algoritmy sú určené čisto len na klasifikáciu (logistická regresia), či regresiu (lineárna regresia) a zas niektoré môžu byť s malými úpravami použité pre oba typy (rozhodovacie stromy, SVM) [10].

Pre náš predikčný nástroj budeme uvažovať o klasifikačných algoritmoch strojového učenia, keďže budeme chcieť predikovať CNV, či je buď patogénne alebo benígne, príp. pravdepodobnosť patogenicity/benignity.

2.2.2 Algoritmy strojového učenia

Algoritmy strojového učenia sú technikami odhadujúcimi funkciu *(f)* pre predikovanie výstupu *(Y)* zo zadaných vstupných údajov *(X)*. Tiež môžeme povedať, že uskutočňujú učenie modelu. Natrénovaný predikčný model potom obsahuje naučené vzťahy a je pripravený predikovať nové vstupy.

Tieto algoritmy sa líšia vo vytvorení funkcie. Keďže nevieme vopred určiť ktorý algoritmus bude pre náš problém najlepší (najlepší v odhadovaní funkcie a následne najlepší v predikovaní), preto vyskúšame rôzne algoritmy strojového učenia, ktoré navzájom porovnáme.

Rozhodovacie stromy

Rozhodovacie stromy predstavujú sadu a postupnosť pravidiel - rozhodovacích kritérií, ktoré sú usporiadané hierarchicky. Ne/Splnenie podmienok (pravidliel) sa vyhodnocuje postupne od koreňa, cez vnútorné uzly, až po listy (koncové uzly), v ktorých sa nachádza výsledná/predikovaná kategória, do ktorej spadá vyhodnocovaná vzorka. [29]

Základným algoritmom na vytváranie binárnych rozhodovacích stromov je algoritmus *CART*. Pri ňom sa na začiatku nachádzajú všetky pozorovania v koreni (počiatoč-

nom uzle). Následne sú rozdelené do dvoch vnútorných uzlov, podľa závislej premennej Y a prediktoru X, a tie sa znovu rozdelia do dvoch uzlov (do vnútorného uzla alebo listu). Hodnoty premennej Y by mali byť čo najviac odlišné od hodnôt v iných uzloch. [29]

Považujú sa za *greedy* algoritmy, keďže zakaždým sa snaží vytvoriť optimálne rozdelenie (podmienku v uzle) [41].

Random forest

Táto metóda združuje viacero rozhodovacích stromov, pričom výsledná predikcia je tvorená spriemerovaním predikcií vytvorených prediktorov (rozhodovacích stromov). Využíva *bootstrap agregáciu* a *random subspace* metódu aby sa zlepšili výsledky predikcií dát na ktorých model nebol trénovaný, čo bol problém, ktorý vzniká v samotných rozhodovacích stromoch. *Bootstrap agregácia (bagging)* je metóda, pri ktorej sa vytvárajú viaceré prediktory (rozhodovacie stromy) a následne tieto sa zagregujú do výsledného prediktora. Náhodnosť je v tom, že pre vytvárajúce sa prediktory sa náhodne vyberú záznamy z trénovacích dát, ktoré budú tvoriť nový trénovací set pre trénovanie prediktoru. Tieto záznamy sa môžu znovu použiť v trénovaniach ďalších prediktorov, keďže sa vracajú späť do pôvodnej trénovacej sady. Výslednou hodnotou predikcie sú spriemerované predikcie všetkých vzniknutých prediktorov. [5], [6], [41]

XGBoost

Optimalizovaná verzia *gradient boosting* algoritmu – XGBoost (*eXtreme Gradient Boosting*), čo bol aj zámer jeho tvorcov vylepšiť ju tak, aby bola vysoko efektívna a flexibilná [11]. To sa im aj podarilo, z hľadiska rýchlosti kvôli paralelným výpočtom (S. Pafka: porovnavací test [37]), vysokej presnoti predikcií (víťaz Kaggle súťaží) a efektívnej spotrebe pamäte (spracovanie signálov z hadrónového urýchľovača v CERNe). [8], [46], [21], [11]

XGBoost implementuje *gradient boosting decision trees* algoritmus a teda základom sú rozhodovacie stromy. Z princípu svojho návrhu je vhodný pre spracovanie tabuľkových dát, preto sme sa ho rozhodli v práci aj tento použiť. Samotný *boosting* je technikou, pri ktorej sa nové modely (stromy) vytvárajú postupne tak, aby každý nasledujúci strom znížil chyby predchádzajúceho. Teda každý sa učí z chýb svojich predchodcov a tým sa snaží znižovať chyby toho predchádzajúceho, ale nezmení predchádzajúci model, iba vznikne nový model, ktorý sa 'poučil' z chýb predchádzajúcich modelov. Ako sa modely postupne vytvárajú, tak aj ich predikcie sa kombinujú, aby sa celková predikcia zvýšila. Základom je slabý model s nízkou presnosťou a následnou kombináciou novovzniknutých modelov získame silný model. Prvou implementáciou boostingu bol *AdaBoost* (Adaptive Boosting). [8], [46], [21], [11], [22]

Pri *gradient boosting*u sú postupne vytvárané nové modely, ktoré predikujú chyby prechádzajúcich modelov. Tiež na začiatku je slabý model a následne sa postupne vytvárajú nové modely, ktoré sa kombinujú. V tejto postupnosti sa najprv vypočítajú chyby pre každé pozorovanie v dátovej sade. Potom sa vytvorí model, ktorý tieto chyby predikuje. Predikcia z tohto modelu sa sčíta s predikciami predchádzajúcich modelov. [8], [3]

Pri *gradient boosting*u sa vzniknuté chyby minimalizujú pomocou *gradient descent* algoritmu, váhovaním podľa gradientu tak, že postupne optimalizuje rozdiel medzi predikovanou a skutočnou hodnotou novým nastavením váh [21].

Problémové a veľmi náročné na pamäť je rozdelovanie v uzloch stromov (nakoľko sa skúša každé jedno rozdelenie a následne sa vyberie najlepšie), zefektívnili tým, že najprv usporiadajú dáta podľa hodnôt atribútov [11].

K ďalším vylepšeniam XGBoostu podľa [46] a [11] partí:

- optimalizácia hĺbky stromov,
- optimalizácia dostupného miesta na disku a maximalizácia jeho využitia pri spracovávaní veľkých súborov dát, ktoré sa nezmestia do pamäte,
- rýchlejšie výpočty vďaka blokovej štruktúre pre paralelné vytváranie stromov,
- práca s chýbajúcimi dátami.

2.2.3 Výber modelu klasifikátora

Niektoré klasifikačné algoritmy, popísané v predchádzajúcej časti, boli použité na natrénovanie modelov na pripravených trénovacích setoch. Funkcia na natrénovanie modelu bola použitá s defaultnými hodnotami parametrov. Všetky modely boli použité pomocou implementovanej knižnice *scikit-learn* [39], okrem XGBoost-u ktorého implementácia je dostupná na [11]. Ich predikcie na trénovacom a validačnom sete sme navzájom porovnali.

2.2.4 Štatistické metódy na vyhodnocovanie správností testov

Na ohodnotenie natrénovaných modelov sme využili nasledovné štatistické metódy pre vyhodnocovanie správností testov. Nasledujúce informácie v tabuľke 2.1 čerpané z [38]. TP = správne vyhodnotené patogénne CNV, TN = správne vyhodnotené benígne CNV, FP = benígne CNV vyhodnotené ako patogénne, FN = patogénne CNV vyhodnotené ako benígne

Presnosť (accuracy)	$\frac{TP+TN}{TP+TN+FP+FN}$	Celková presnosť predstavuje podiel správne vyhodnotených CNV ku všetkým hodnoteným.
Pozitívna predikčná hodnota PPV (precision)	$\frac{TP}{TP+FP}$	určuje pravdepodobnosť, že CNV je naozaj pozitívne, keď metóda predikovala, že je CNV pozitívne
Senzitivita	$\frac{TP}{TP+FN}$	je schopnosť predikčnej metódy určiť patogénne CNV ako patogénne (miera úspešnosti detekcie skutočne pozitívnych CNV)
Špecificita	$\frac{TN}{TN+FP}$	je schopnosť predikčnej metódy identifikovať benígne CNV ako benígne (miera úspešnosti detekcie skutočne benígnych CNV)

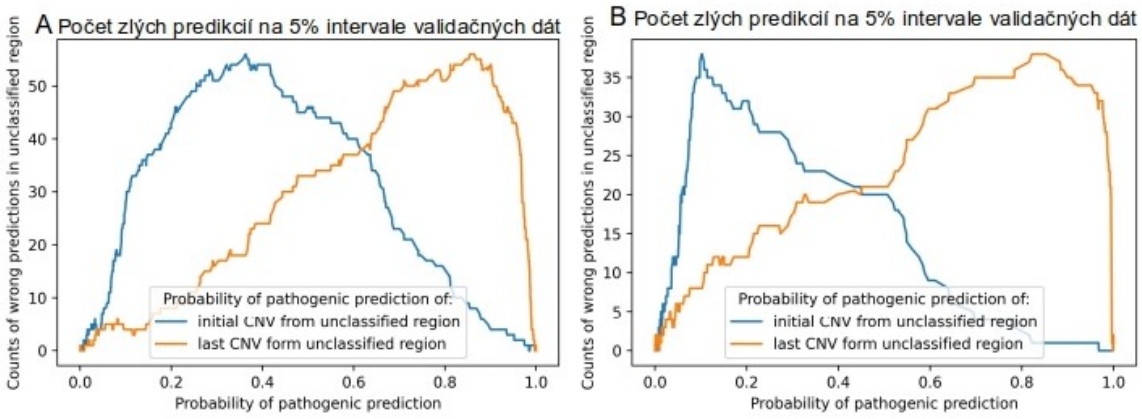
Tabuľka 2.1: Štatistické metódy na vyhodnocovanie správností testov

2.2.5 Identifikácia hraníc pre klasifikovanie CNV ako *neklasifikované*

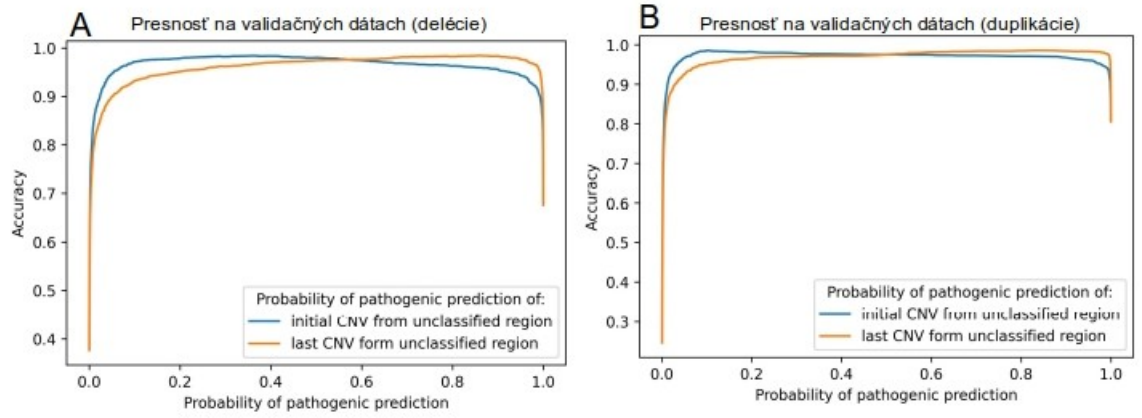
Vo všeobecnosti, pravdepodobnosti okolo 50% sú viac náchylné na nesprávnu klasifikáciu ako pri hraničných hodnotách 0% a 100%. Avšak predikcie patogenicity CNV a následná klasifikácia by mali byť dôveryhodné vzhľadom na citlivosť diagnostických analýz. Tieto výsledky sú veľmi citlivé, keďže chyba vo vyhodnotení má veľmi vážne dopady ako je psychické vypätie pacientov, matiek, či blízkych. Pri nesprávnom vyhodnotení benígneho nálezu podstupujú zbytočné zákroky a naopak pri nesprávnom vyhodnotení patogénneho nálezu to môže vyústiť do podcenenej zdravotnej starostlivosti. Podľa uvedeného je dôležité určiť také hranice, pri ktorých môžeme dôverovať výsledkom. Na základe toho, CNV, ktorých predikcie spadajú do hranice okolo 50% by mali byť klasifikované ako *neklasifikované*. Tieto hranice a teda podiel nálezov, ktoré sme ochotní hodnotiť ako neklasifikované, ako aj výslednú presnosť ktorú chceme dosiahnuť, si vieme zadefinovať ľubovoľne. My sme si to stanovili ako 5 a 10%-ný interval.

Identifikovali sme tieto hranice nasledovným postupom:

V prvom kroku spracované dáta z validačného setu boli zoradené podľa predikovanej pravdepodobnosti patogenicity. Získali sme teda usporiadanú postupnosť CNV z benígnej na patogénnu predikciu. Ďalej sme zrátali koľko nesprávne predikovaných CNV sa nachádzalo v každom 5%-nom úseku posúvaním tohto 5%-ného intervalu (interval predstavuje 126 delécií a 90 duplikácií z testovacej sady) (obr. 2.3). Následne sme zisťovali ako sa zmenila presnosť na 'novej' dátovej sade po odstránení tohto 5%-ného intervalu (obr. 2.4). Presnosť predikcie bola naozaj najvyššia, keď sa odstránil interval s patogénnymi predikciami okolo 50% (obr. 2.4). Nakoniec sme zisťovali pravdepodobnosť patogénnych predikcií odstránených CNV (na začiatku a na konci daného



Obr. 2.3: Počet zlých predikcií na 5% intervale validačných dát
Počet zlých predikcií na 5% intervale validačných dát (A - delécie, B - duplikácie).



Obr. 2.4: Presnosť predikcií po vyhodení 5%-ných intervalov
Presnosť predikcií na validačnom sete po odstránení 5%-ného intervalu, na ktorom sa nachádzalo najviac nesprávne klasifikovaných CNV (A - delécie, B - duplikácie).

intervalu), aby sme stanovili hranice patogénnych predikcií pre klasifikáciu CNV ako *neklasifikované*.

Podobný postup sme použili so stanovením 10%-ného intervalu.

Kapitola 3

Výsledky

V tejto kapitole budú prezentované dosiahnuté výsledky. Všetky kódy, dáta a vygenerované grafy sú dostupné cez verejný GitHub repozitár: `https://github.com/MiskaGaziova/bakalarka_isv`.

3.1 Výber atribútov

Dôležitou súčasťou pri odhaľovaní problému pri výskyte CNV bolo nájsť významné informácie o génoch. V tejto podkapitole si povieme o významných atribútoch génov, ktoré sa nám podarilo vytvoriť zo získaných anotácií nástroja AnnotSV a teda z databáz spomenutých v časti 1.5.

Dĺžka štrukturálnych variantov Jedným z hlavných činiteľov ovplyvňujúcich patogenicitu (čo sa aj fenotypovo môže prejaviť) je dĺžka CNV. Je to spôsobené aj tým, že čím je CNV dlhšie, tak tým väčšie množstvo génov (ako aj promotórov a iných regulačných oblastí) môže zasiahnuť.

Porovnaním dĺžok medzi patogénnymi a benígnymi CNV sme zistili, že patogénne duplikácie a delécie sú ozaj dlhšie ako benígne duplikácie a delécie. Tento výsledok je pozororovateľný na grafe 2.1.

Počet génov nachádzajúcich sa na úseku CNV S dĺžkou variantu je spojený aj počet génov, ktoré CNV zasahuje. To môže byť nežiadúce z hľadiska delécie, pričom môže potrebný gén úplne vymiznúť alebo kóduje len malú čiastku z toho génu, teda môže nasyntetizovaný proteín plniť inú funkciu ako ten pôvodný. Naopak pri duplikácii môže nadmerná expresia génu byť škodlivá pre organizmus.

"Morbídne"gény prekrývajúce úsek CNV Súčasťou OMIM-u [20] sú aj tzv. 'morbídne' gény. Pri mutácii týchto génov môžu narušiť správny chod organizmu a

tým môžu spôsobiť určité ochorenie, teda pri mutácií týchto génov na už jednej kópii vzniká problém.

Pri detekcii tohto 'morbídneho' génu v OMIM, môžu nastať dve situácie a to: molekulárna podstata ochorenia je známa alebo sa stalo to, že syndróm sa prejaví po delécii alebo duplikácii spolu so susednými génmi.

Konzervovanosť génov prekrytých daným CNV Gény, ktoré sú častejšie zastúpené v populácii, by mali byť menej náchylné na zmeny. Preto pri zmene génu, či výskytu množstva génu spôsobujú vážne problémy. Zo všetkých génov, ktoré CNV prekrýva, sme vybrali najnižšiu frekvenciu pre gén, keďže sme spozorovali, že gény nachádzajúce sa v patogénnych CNV majú nižšie zastúpenie v populácii. Populačné frekvencie génov boli získané z databázy DGV [34], čo sú agregované genómy z rôznych štúdií.

Skóre netolerancie straty funkcie pre gén Skóre vypočítané v ExAC databáze [33] tzv. pLI *(probability of being loss-of-function intolerant)* predstavuje pravdepodobnosť, že gén netoleruje stratu funkcie pri mutácii daného génu (*LoF* – loss of function). Pri výpočte ich rozdelili do troch kategórií: 1. *Null* – vtedy gén toleruje stratu funkcie; 2. *Recessive* – pričom gény v heterozygótnom stave tolerujú stratu funkcie; 3. *Haploinsufficient* – tu gény aj v heterozygótnom stave vôbec netolerujú ich stratu a teda môže dôjsť k závažnej zmene fenotypu organizmu. Gény s vysokým skóre pLI (pLI $\geq$ 0,9), ktoré patria do tretej kategórie, sú extrémne netolerantné na LoF, zatiaľ čo gény s nízkym skóre pLI (pLI $\leq$ 0,1) tolerujú LoF. Anotáciu pre CNV sme vytvorili počtom génov, ktorých toto skóre bolo väčšie. [33], [49]

ExAC Z-skóre génu Databáza ExAC [33] poskytuje tiež Z-skóre nefunkčnosti génu v dôsledku delécie alebo duplikácie nachádzajúcej sa na úseku CNV. Vzhľadom k tomu, že čím vyššie je toto skóre, tým je to problematickejšie pre organizmus, ak je tento gén deletovaný alebo duplikovaný, preto z tohto skóre sme pre CNV vybrali to maximálne zo všetkých skóre pre gény nachádzajúce sa na úseku daného CNV.

Skóre haploinsuficiencie z DDD Databáza DDD [15] poskytuje skóre haploinsuficiencie pri ktorom nastáva, že jedna funkčná kópia génu je nedostatočná na udržanie normálnej funkcie génového produktu. Anotáciu pre CNV (iba delécie) z tohto atribútu sme vytvorili počtom génov s týmto skóre menším ako 10% nakoľko menšie skóre je problematickejšie.

Skóre haploinsufficiencie a triplosensitivity z ClinGen-u Aj ClinGen databáza [43] (ClinGen consortium rating system) obsahuje skóre (0, 1, 2, 3) haploinsufficiencie

a triplosensitivity génov. Skóre 0 – zatiaľ žiadny dôkaz o prítomnosti patogenicity pri duplikácii alebo delécii, ..., skóre 3 – dôkazy o prejave patogenicity pri výskyte duplikácii alebo delécii daného génu sú známe. Anotáciu pre CNV sme vytvorili počtom génov s vysokým (*2* a *3*) skóre haploinsufficiencie (triplosensitivity) pre delécie (duplikácie).

Výsledný výber a použitie atribútov v modeli je znázornené v tabuľke 3.1. Pri selektovaní atribútov sme nakoniec nevybrali atribút počet promótorov, keďže bol totožný s počtom génov prekrytých CNV. Rozhodli sme sa tak spraviť, v dôsledku toho, že údaj o lokalizácii promotóru bol fixná pozícia – 500 bp upstream od začiatku transkripcie (pozície začiatkov transkripcie boli získané z databáz RefSeq a ENSEMBL).

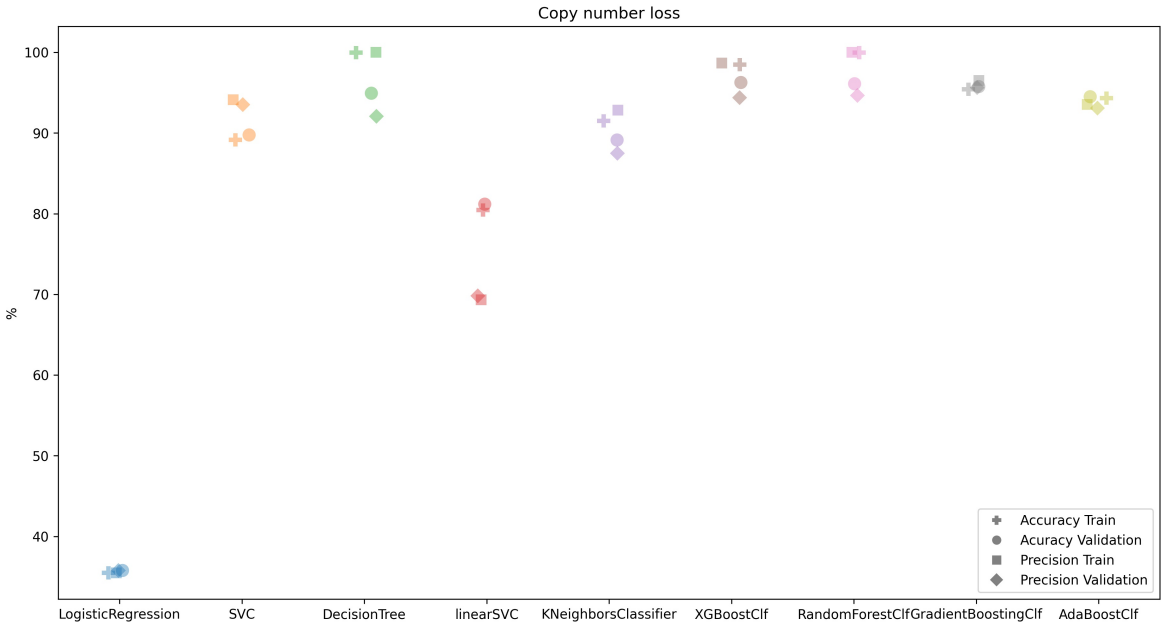
Názov atribútu	Delécie	Duplikácie
Dĺžka CNV	áno	áno
Počet génov, ktoré CNV prekrýva	áno	áno
Počet 'morbídnych' génov, ktoré CNV prekrýva	áno	áno
Konzervovanosť génov prekrytých daným CNV	áno	áno
Skóre netolerancie straty funkcie pre gén (ExACzScore pLI)	áno	áno
ExAC Z-skóre génu (ExACzScore indel)	áno	áno
Skóre haploinsuficiencie z DDD	áno	nie
Skóre haploinsufficiencie a triplosensitivity z ClinGen-u	áno	nie

Tabuľka 3.1: Tabuľka znázorňuje (ne)použitie atribútov v modeloch pre duplikácie a delécie

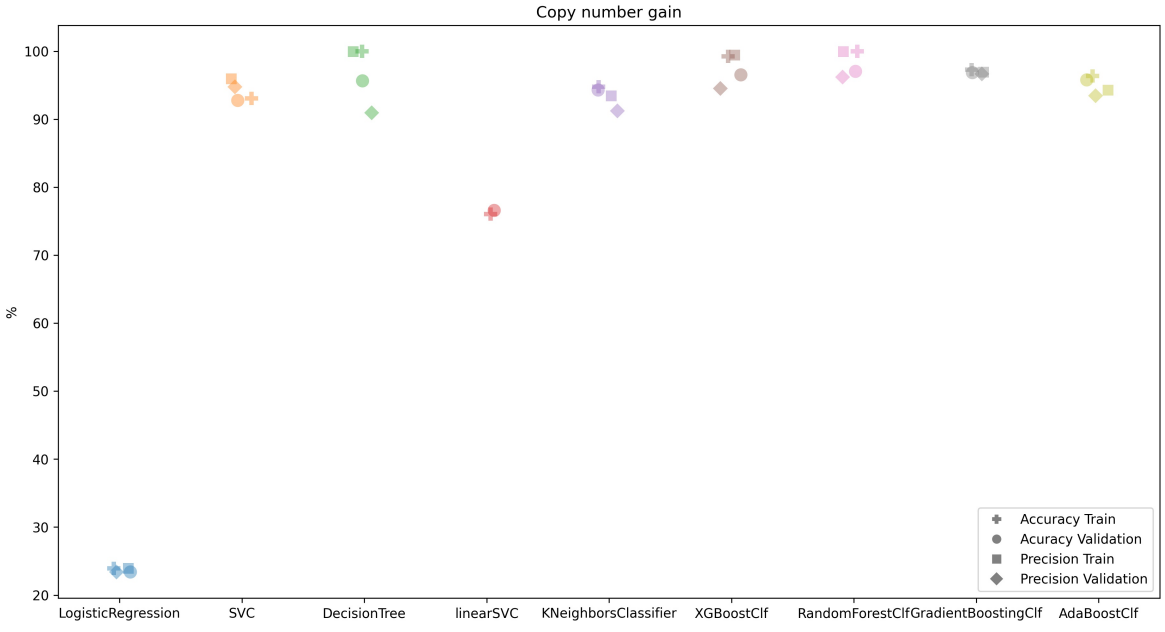
3.2 Výber modelu

Porovnali sme presnosti niekoľkých predikčných metód na trénovacích a validačných dátach. Presnosti modelov na oboch dátových sadách sú znázornené na obr. 3.1 pre delécie a obr. 3.2 pre modely trénované na duplikáciách. Vyhodnotenia všetkých modelov sú k nahliadnutiu v tabuľkách v prílohe pre delécie tab. 3.5 a duplikácie tab. 3.6. Zistili sme, že natrénované modely pomocou XGBoost-u a RandomForest-u sú veľmi porovnateľné a na pripravených dátach vykazujú najlepšie správanie. Pre nastávajúcu prácu sme si vybrali model XGBoost-u.

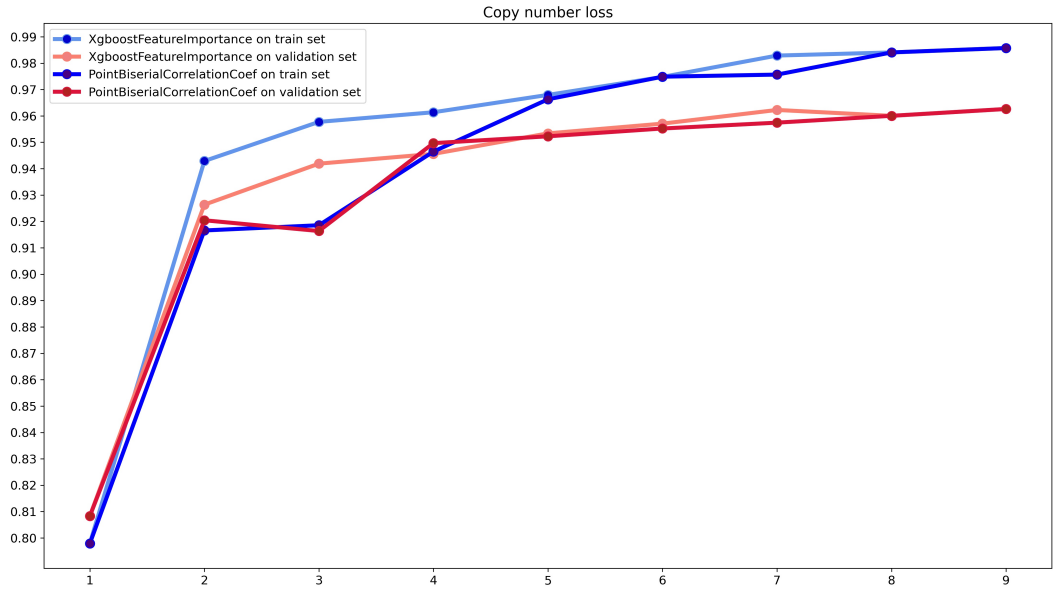
Zisťovali sme, ako sa správa presnosť predikčného modelu pridávaním atribútov podľa Point-Biserial korelačného koeficientu a podľa metódy feature_importance implementovanej v XGBoost knižnici, kde každý z vybraných atribútov prispieval ku klasifikácii. Vizualizácie správania sa presností modelov postupným pridávaním atribútov sú k nahliahnutu na grafoch: pre delécie obr. 3.3 a duplikácie obr. 3.4. Názvy



Obr. 3.1: Porovnanie presností predikčných metód - delécie
Porovnanie presností predikčných metód na trénovacích a validačných dátach.



Obr. 3.2: Porovnanie presností predikčných metód - duplikácie
Porovnanie presností predikčných metód na trénovacích a validačných dátach.



Obr. 3.3: Vizualizácia zvyšovania presnosti modelu na základe postupného pridávania atribútov – delécie

atribútov pre čísla v grafoch sú v tabuľkách v prílohe (tab. 3.7 a 3.8).

Pozreli sme sa aj ako model využíva pri trénovaní vybrané atribúty - akú dôležitosť majú dané atribúty pre trénovanie modelu (obr. 3.5).

3.3 Určenie hraníc pre klasifikáciu CNV ako *neklasifikované*

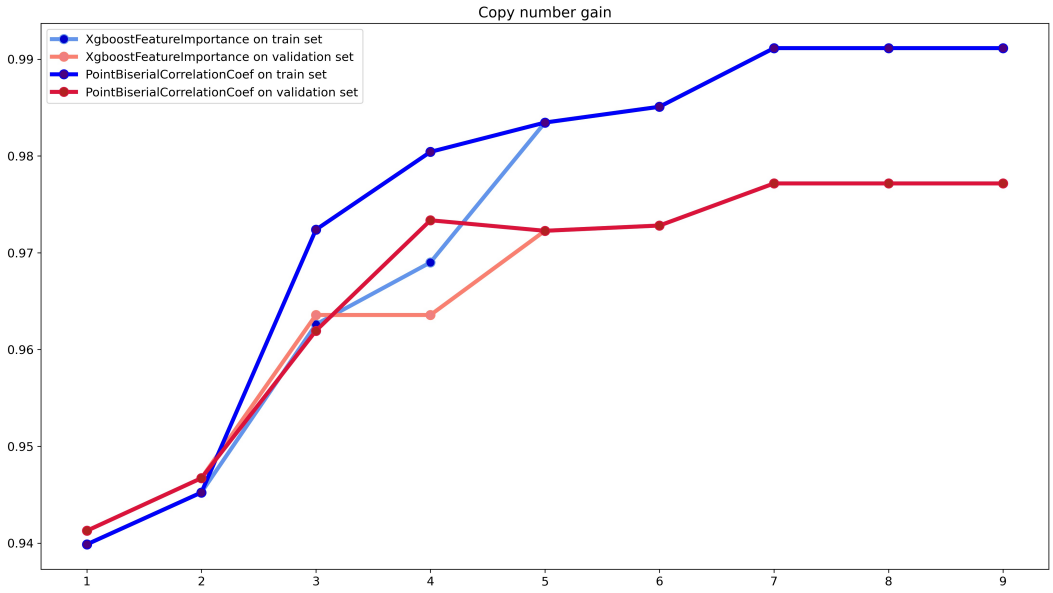
Hľadali sme hranice predikovanej pravdepodobnosti patogenicity, ktorú by sme mohli klasifikovať ako *neklasifikované*. Stanovením určitého rozmedzia (5 a 10 %-ný interval) a vyhodením tohto úseku, v ktorom sa nachádza najviac zlých predikcií (čím sa zvýši presnosť), sme zistili nasledovné:

Ak sme vyhodili 5% dát z validačného setu:

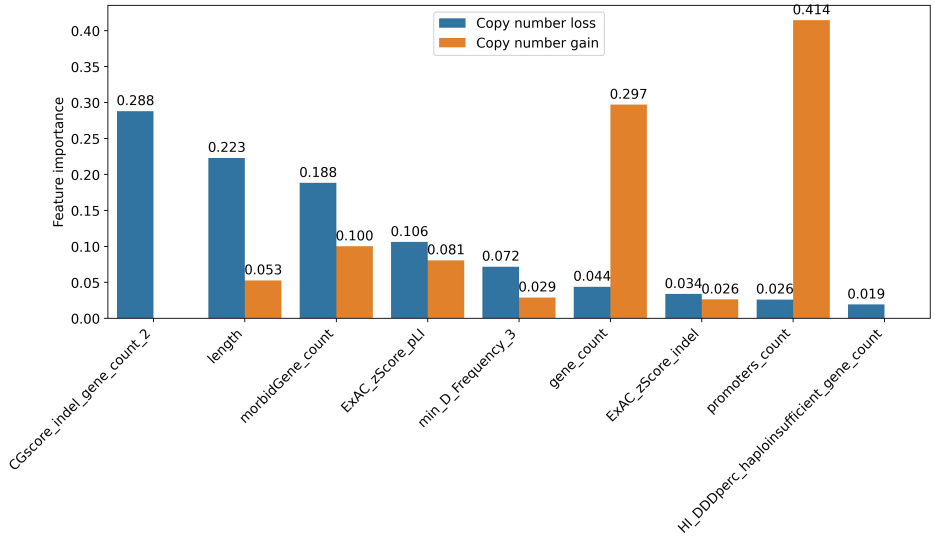
- Hranice pravdepodobností patogenicity u delécií: 35% – 85% by sme mohli klasifikovať ako neklasifikované, keďže zlých predikcií bolo na tomto úseku 56 delécií (zo 126) –> presnosť na validačnom sete by sa zvýšila na 98,4%

- Hranice pravdepodobností patogenicity pre duplikácie: 10% – 85% by sme mohli klasifikovať ako neklasifikované, keďže zlých predikcií bolo na tomto úseku 38 duplikácii (z 90) –> presnosť na validačnom sete by sa zvýšila na 98,6%

Ak sme vyhodili 10% dát z validačného setu:



Obr. 3.4: Vizualizácia zvyšovania presnosti modelu na základe postupného pridávania atribútov – duplikácie



Obr. 3.5: Dôležitosť atribútov pre trénovanie XGBoost modelu

- Hranice pravdepodobností patogenicity u delécií: 11% – 93% by sme mohli klasifikovať ako neklasifikované, keďže zle klasifikovaných bolo na tomto úseku 79 CNV (z 252) –> presnosť na validačnom sete by sa zvýšila na 99,3%
- Hranice pravdepodobností patogenicity pre duplikácie: 4% – 94% by sme mohli klasifikovať ako neklasifikované, keďže zle klasifikovaných bolo na tomto úseku 48 CNV (zo 180) –> presnosť na validačnom sete by sa zvýšila na 99%

3.4 Presnosti výsledného modelu

Po natrénovaní **model** vykazoval presnosti na testovacích dátach na deléciách 96% a na duplikáciách 97.3%. Ako spadajú predikované CNV do jednotlivých kategórií (správne a nesprávne predikcie patogénnych a benígnych CNV) je zvizualizované na obr. 3.6. Presnosti výsledného modelu pre všetky dátové sady (trénovací, validačný a testovací) sú v tab. 3.2.

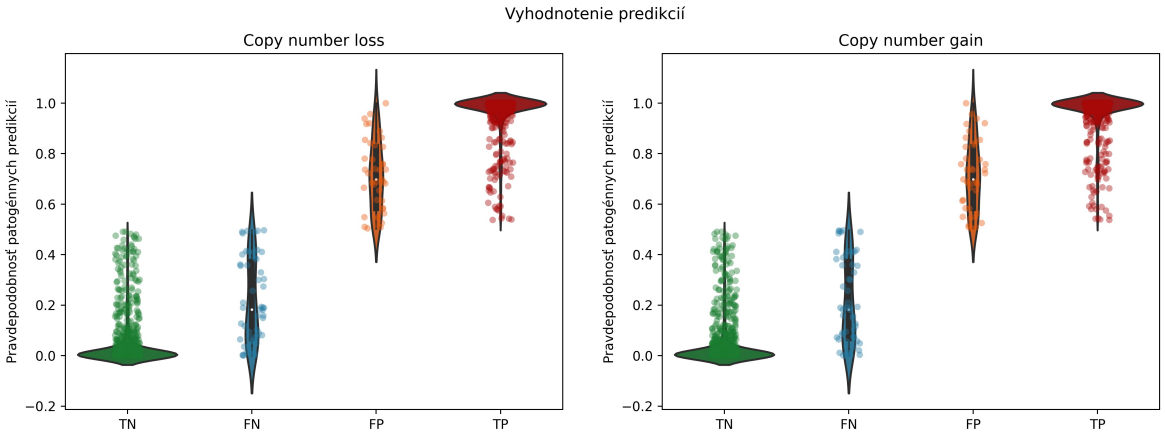
CNV	Dataset	Presnosť	PPV	Sensitivita	Špecificita
del	train	98.57%	98.47%	97.26%	99.24%
del	val	96.26%	95.60%	92.91%	97.90%
del	test	96.04%	95.76%	91.97%	98.02%
dup	train	99.11%	99.50%	96.77%	99.85%
dup	val	97.72%	96.33%	93.59%	98.94%
dup	test	97.34%	96.36%	91.38%	99.02%

Tabuľka 3.2: Výsledné presnosti modelov pre delécie a duplikácie
Výsledné presnosti modelov pre duplikácie (dup) a delécie (del) na trénovacom (train), validačnom (val) a testovacom (test) sete.

Model so stanovenými hranicami pre klasifikáciu CNV ako *neklasifikované*
Po stanovení hraníc, ktorými klasifikujeme CNV ako neklasifikované, ak predikcia spadá do tohto intervalu, sa presnosti zvýšili aj na testovacom sete (viď tab. 3.3). Vizualizácia vyhodnotenia na testovacom sete klasifikovaním aj neklasifikovaných je znázornená na obr. 3.7 a 3.8. Rozhodli sme sa ďalej využívať hranice stanovené 5%-ným intervalom.

3.5 Porovnanie s dostupnými klasifikačnými nástrojmi

Spravili sme aj porovnania s niektorými dostupnými klasifikačnými nástrojmi (spomenutými v časti 1.3) ako oni vyhodnotia a klasifikujú CNV z pripraveného testovacieho setu.

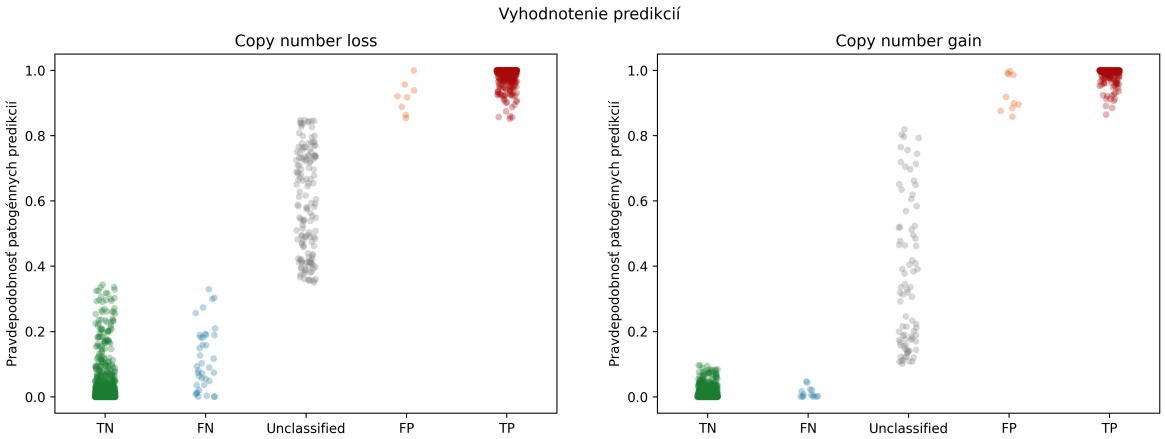


Obr. 3.6: Výsledné predikcie delécií a duplikácií

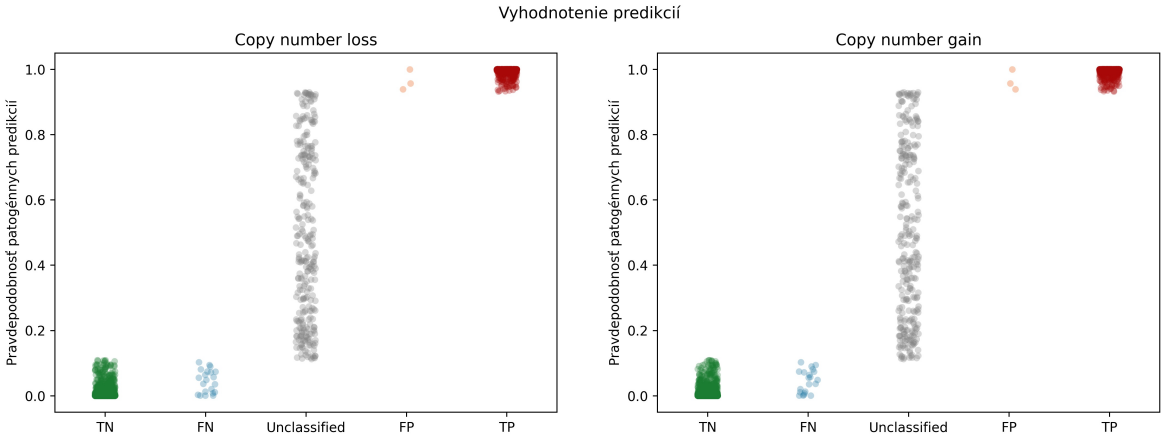
Graf znázorňuje správne a nesprávne predikovanie CNV z testovacieho setu (delécií - copy number loss a duplikácií - copy number gain). *TP* = správne vyhodnotené patogénne CNV, *TN* = správne vyhodnotené benígne CNV, *FP* = nesprávne vyhodnotené benígne CNV, *FN* = nesprávne vyhodnotené patogénne CNV

	Presnosť	
CNV	**5%-ný interval neklasifikovaných CNV**	**10%-ný interval neklasifikovaných CNV**
del	98.06%	98.93%
dup	98.61%	99.02%

Tabuľka 3.3: Presnoti modelu s použitím hraníc na klasifikáciu neinformatívnych Presnoti modelu na testovacích dátach s vyhodením CNV, ktorých predikcie spadali do stanovených hraníc klasifikácie neinformatívnych.



Obr. 3.7: Vizualizácia predikcií delécií a duplikácií aplikovaním klasifikácie *neklasifikovaných CNV* s využitím stanovenej hranice pravdepodonosti patogenicity podľa 5%-ného intervalu



Obr. 3.8: Vizualizácia predikcií delécií a duplikácií aplikovaním klasifikácie *neklasifikovaných CNV* s využitím stanovenej hranice pravdepodonosti patogenicity podľa 10%-ného intervalu

CNV	**Presnosť**	**FP**	**TN**	**TP**	**FN**	**nejasný význam**
del	98.725%	9	229	547	1	1740
dup	99.38%	3	196	285	0	1324

Tabuľka 3.4: ClassifyCNV vyhodnotenie na testovacom sete

CNV z testovacieho setu boli ohodnotené nástrojom AnnotSV – nakoľko všetky CNV boli anotované týmto nástrojom, tak aj výsledná klasifikácia nástroja AnnotSV bola súčasťou anotácií. Tu presnosť dosahovala 73% na deléciách a 65% na duplikáciách, pričom žiadne z týchto CNV nebolo klasifikované ako *nejasný* význam.

Ďalším dostupným nástrojom bol ClassifyCNV [19], ten dosahuje vysoké presnosti, ale veľkú časť klasifikuje ako *nejasný* význam. ClassifyCNV dosahoval vysokú úspešnosť na testovacom sete – a to až 98.7% na deléciách a 98.4% na duplikáciách, avšak až 68% z delécií (1740 CNV) a 73% z duplikácií (1324 CNV) bolo vyhodnotených ako *nejasný* význam (viď tab. 3.4).

Prezentovaná metóda pri použití stanovených hraníc (v časti 3.3) na klasifikovanie ako neklasifikované na testovacích dátach vyhodnotila iba 5% zo vzoriek CNV ako neklasifikované.

Záver

Cieľom bakalárskej práce bolo vyhľadať významné atribúty popisujúce CNV a vytvoriť predikčnú metódu na určenie klinického dopadu ľudských CNV. K tomuto účelu sme zadefinovali dôležité biologické pojmy, spomenuli možné následky prítomnosti CNV, vyhľadali dostupné anotačné a klasifikačné nástroje a zdroje. Nakoniec sme popísali štatistické metódy na porovnávanie atribútov a vybrané algoritmy strojového učenia.

Na začiatku sme si zozbierali CNV nálezy so známym klinickým významom z verejne dostupných genomických databáz (v časti 2.1) a vytvorili dátové sady pre modelovanie. Nálezy boli anotované atribútmi, z ktorých sme vybrali menšiu skupinu významných atribútov. Vybrali sme tým vhodné atribúty pre trénovanie modelov (časť 3.1). Porovnali sme presnosti modelov na pripravených dátach v časti 3.2. Zistili sme, že na pripravených dátach modely XGBoost-u a RandomForestu vykazujú podobné správanie. Pre natrénovaný model na XGBoost-e sme stanovili hranice pravdepodobnosti patogenicity pre klasifikáciu CNV ako neklasifikovaných (časť 3.3). Pre delécie sme určili 35% – 85% a pre duplikácie 10% – 85% hranice pravdepodobnosti patogenicity pre klasifikovanie CNV ako *neklasifikovaných*. Na testovacích dátach modely vykazovali vysoké presnosti – na deléciách 96% a na duplikáciách 97.3% (presnosti pre všetky dátové sady v tab. 3.2). Po stanovení hraníc pravdepodobnosti patogenicity pre neklasifikované sa zvýšili presnosti na 98% u delécií a na 98.6% u duplikácií. Porovnali sme výsledky predikcii metódy s dostupnými nástrojmi (ClassifyCNV a AnnotSV). Zistili sme, že vytvorená metóda je presnejšia v klasifikácii v porovnaní s AnnotSV a v porovnaní s ClassifyCNV je v počte klasifikovaných CNV ako *nejasný význam* lepšia, keďže klasifikovala oveľa menej CNV ako *neklasifikované*.

Na práci sa dá pokračovať pridaním nových genomických atribútov z ďalších genomických databáz. Ďalším krokom vylepšenia metódy a zvýšenia jej presnosti predikcií je nastavenie vhodných parametrov modelu (napr. pomocou grid search algoritmu či cross validácie). Vhodné by bolo zistiť, ako sa model správa na CNV ohodnotených ako *likely pathogenic* a *likely benign* v ClinVar-e; poprípade vytvoriť nový model, ktorý by bol natrénovaný len na takto hodnotených CNV. A nakoniec vyhodnotiť na nezávislých vzorkách – CNV také, ktoré sa nenachádzajú v pripravených dátových sadách (ClinVar databáze), ako napríklad CNV spôsobujúce závažné syndrómy a ochorenia z

OMIM, či DECIPHER databázy alebo benígne CNV s vysokou populačnou frekvenciou z gnomAD databázy.

Literatúra

[1] B Alberts, A Johnson, J Lewis, M Raff, K Roberts, and P Walter. Molecular Biology of the Cell 5th edition. *NY: Garland Science*, 2007.

[2] Joanna S Amberger, Carol A Bocchini, Alan F Scott, and Ada Hamosh. omim.org: leveraging knowledge across phenotype–gene relationships. *Nucleic acids research*, 47(D1):D1038–D1043, 2019.

[3] Dans Becker. XGBoost. `https://www.kaggle.com/dansbecker/xgboost`, 2018. Kaggle.

[4] Yoshua Bengio, Ian Goodfellow, and Aaron Courville. *Deep learning*, volume 1. MIT press Massachusetts, USA:, 2017.

[5] Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

[6] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[7] James Dean Brown. Point-biserial correlation coefficients. *Statistics*, 5(3), 2001.

[8] Jason Brownlee. A gentle introduction to XGBoost for applied machine learning. `https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/`. Machine Learning Mastery.

[9] Jason Brownlee. *Master Machine Learning Algorithms: discover how they work and implement them from scratch.* Machine Learning Mastery, 2016.

[10] Jason Brownlee. Difference between classification and regression in machine learning. *Machine Learning Mastery*, 25, 2019.

[11] 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*, pages 785–794. ACM, 2016.

[12] Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, and Kent WJ. USCS Genome - Table Browser, 2019. [Citované 2019-4-20] Dostupné z `https://genome.ucsc.edu/cgi-bin/hgTables`.

[13] Teresa Davoli, Andrew Wei Xu, Kristen E Mengwasser, Laura M Sack, John C Yoon, Peter J Park, and Stephen J Elledge. Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. *Cell*, 155(4):948–962, 2013.

[14] Helen V Firth, Shola M Richards, A Paul Bevan, Stephen Clayton, Manuel Corpas, Diana Rajan, Steven Van Vooren, Yves Moreau, Roger M Pettett, and Nigel P Carter. DECIPHER: database of chromosomal imbalance and phenotype in humans using ensembl resources. *The American Journal of Human Genetics*, 84(4):524–533, 2009.

[15] Helen V Firth, Caroline F Wright, and DDD study. The deciphering developmental disorders (DDD) study. *Developmental Medicine & Child Neurology*, 53(8):702–703, 2011.

[16] Véronique Geoffroy, Yvan Herenger, Arnaud Kress, Corinne Stoetzel, Amélie Piton, Hélène Dollfus, and Jean Muller. AnnotSV: an integrated tool for structural variations annotation. *Bioinformatics*, 34(20):3572–3574, 2018.

[17] Véronique Geoffroy, Yvan Herenger, Arnaud Kress, Corinne Stoetzel, Amélie Piton, Hélène Dollfus, and Jean Muller. Annotsv: runjob, 2018. [Citované 2020-5-27] Dostupné z `https://lbgi.fr/AnnotSV/runjob`.

[18] Tatiana A Gurbich and Valery Vladimirovich Ilinsky. ClassifyCNV. `https://github.com/Genotek/ClassifyCNV`, 2020.

[19] Tatiana A Gurbich and Valery Vladimirovich Ilinsky. ClassifyCNV: a tool for clinical annotation of copy-number variants. *Scientific reports*, 10(1):1–7, 2020.

[20] Ada Hamosh, Alan F Scott, Joanna Amberger, David Valle, and Victor A McKusick. Online Mendelian Inheritance in Man (OMIM). *Human mutation*, 15(1):57–61, 2000.

[21] Rohan Harode. Xgboost: A Deep dive into Boosting, Feb 2020.

[22] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

[23] Taobo Hu, Si Chen, Ata Ullah, and Hong Xue. Aluscancnv2. `https://github.com/hutaobo/AluScanCNV2`, 2019.

[24] Taobo Hu, Si Chen, Ata Ullah, and Hong Xue. AluScanCNV2: An R package for copy number variation calling and cancer risk prediction with next-generation sequencing data. *Genes & diseases*, 6(1):43–46, 2019.

[25] National Cancer Institute. NCI Dictionary of genetics terms. `https://www.cancer.gov/publications/dictionaries/genetics-dictionary/def/haploinsufficiency`.

[26] Baltimore Johns Hopkins University. Online Mendelian Inheritance in Man, OMIM. `https://www.omim.org/entry/601410?search=601410&highlight=601410`, 2019. MIM Number: 601410.

[27] Eric Jones, Travis Oliphant, and Pearu Peterson. SciPy: Open source scientific tools for Python. 01 2001.

[28] Konrad J Karczewski, Laurent C Francioli, Grace Tiao, Beryl B Cummings, Jessica Alföldi, Qingbo Wang, Ryan L Collins, Kristen M Laricchia, Andrea Ganna, Daniel P Birnbaum, et al. Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *BioRxiv*, page 531210, 2019.

[29] Klára Komprdová et al. *Rozhodovací stromy a lesy*. Akademické nakladatelství CERM, 2012.

[30] Max Kuhn, Kjell Johnson, et al. *Applied predictive modeling*, volume 26. Springer, 2013.

[31] Melissa J. Landrum, Jennifer M. Lee, Mark Benson, Garth Brown, Chen Chao, Shanmuga Chitipiralla, Baoshan Gu, Jennifer Hart, Douglas Hoffman, Jeffrey Hoover, Wonhee Jang, Kenneth Katz, Michael Ovetsky, George Riley, Amanjeev Sethi, Ray Tully, Ricardo Villamarin-Salomon, Wendy Rubinstein, and Donna R. Maglott. ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Research*, 44(D1):D862–D868, 11 2015.

[32] Ryan M Layer, Colby Chiang, Aaron R Quinlan, and Ira M Hall. LUMPY: a probabilistic framework for structural variant discovery. *Genome biology*, 15(6):R84, 2014.

[33] Monkol Lek, Konrad J Karczewski, Eric V Minikel, Kaitlin E Samocha, Eric Banks, Timothy Fennell, Anne H O'Donnell-Luria, James S Ware, Andrew J Hill, Beryl B Cummings, et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 536(7616):285–291, 2016.

[34] Jeffrey R MacDonald, Robert Ziman, Ryan KC Yuen, Lars Feuk, and Stephen W Scherer. The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic acids research*, 42(D1):D986–D992, 2014.

[35] Patrick E McKnight and Julius Najab. Mann-whitney u test. *The Corsini encyclopedia of psychology*, pages 1–1, 2010.

[36] Beata Nowakowska. Clinical interpretation of copy number variants in the human genome. *Journal of applied genetics*, 58(4):449–457, 2017.

[37] Szilard Pafka. Benchmarking random forest implementations. `https://github.com/szilard/benchm-ml`, 2019.

[38] M Pagano and K Gauvreau. *Principles of biostatistics*. Cengage Learning, 2nd edition, 2000.

[39] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[40] Ondrej Pös, Jan Radvanszky, Jakub Styk, Zuzana Pös, Gergely Buglyó, Michal Kajsik, Jaroslav Budis, Bálint Nagy, and Tomas Szemes. Copy Number Variation: Methods and Clinical Applications. *Applied Sciences*, 11(2):819, 2021.

[41] Kurtis Pykes. Random Forest overview. `https://towardsdatascience.com/random-forest-overview-746e7983316`, Mar 2020.

[42] Richard Redon, Shumpei Ishikawa, Karen R Fitch, Lars Feuk, George H Perry, T Daniel Andrews, Heike Fiegler, Michael H Shapero, Andrew R Carson, Wenwei Chen, et al. Global variation in copy number in the human genome. *nature*, 444(7118):444–454, 2006.

[43] Heidi L Rehm, Jonathan S Berg, Lisa D Brooks, Carlos D Bustamante, James P Evans, Melissa J Landrum, David H Ledbetter, Donna R Maglott, Christa Lese Martin, Robert L Nussbaum, et al. ClinGen – the clinical genome resource. *New England Journal of Medicine*, 372(23):2235–2242, 2015.

[44] Erin Rooney Riggs, Erica F Andersen, Athena M Cherry, Sibel Kantarci, Hutton Kearney, Ankita Patel, Gordana Raca, Deborah I Ritter, Sarah T South, Erik C Thorland, et al. Technical standards for the interpretation and reporting of constitutional copy-number variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics (ACMG) and the Clinical Genome Resource (ClinGen). *Genetics in Medicine*, 22(2):245–257, 2020.

[45] Marián Rimarčík. *Štatistika pre prax*. Štatistika pre prax, 2007.

[46] Ramya Bhaskar Sundaram. An end-to-end guide to understand the math behind XGBoost. `https://www.analyticsvidhya.com/blog/2018/09/an-end-to-end-guide-to-understand-the-math-behind-xgboost/`, Sep 2018.

[47] Michael Waskom. Seaborn - python library. [Citované 2019-10-24] Dostupné z `https://seaborn.pydata.org/examples/index.html`.

[48] Wes McKinney. Data Structures for Statistical Computing in Python. In *Proceedings of the 9th Python in Science Conference*, pages 56 – 61, 2010.

[49] Alban Ziegler, Estelle Colin, David Goudenège, and Dominique Bonneau. A snapshot of some pLI score pitfalls. *Human mutation*, 2019.

Príloha A

Data	**Model**	**TN**	**FP**	**FN**	**TP**	**Presnosť**	**PPV**	**Senzitivita**	**Špecificita**
train	LogisticRegres	0	7608	0	4180	0.35	0.35	1.00	0.00
val	LogisticRegres	0	1622	0	904	0.36	0.36	1.00	0.00
test	LogisticRegres	0	1635	0	891	0.35	0.35	1.00	0.00
train	SVC	7415	193	1084	3096	0.89	0.94	0.74	0.97
val	SVC	1574	48	210	694	0.90	0.94	0.77	0.97
test	SVC	1581	54	229	662	0.89	0.92	0.74	0.97
train	linearSVC	7608	0	4180	0	0.65		0.00	1.00
val	linearSVC	1622	0	904	0	0.64		0.00	1.00
test	linearSVC	1635	0	891	0	0.65		0.00	1.00
train	KneighborsClf	7342	266	737	3443	0.91	0.93	0.82	0.97
val	KneighborsClf	1517	105	169	735	0.89	0.88	0.81	0.94
test	KneighborsClf	1535	100	182	709	0.89	0.88	0.80	0.94
train	XGBoostClf	7553	55	124	4056	0.98	0.99	0.97	0.99
val	XGBoostClf	1571	51	43	861	0.96	0.94	0.95	0.97
test	XGBoostClf	1594	41	55	836	0.96	0.95	0.94	0.97
train	RandomForest	7608	0	2	4178	1.00	1.00	1.00	1.00
val	RandomForest	1574	48	53	851	0.96	0.95	0.94	0.97
test	RandomForest	1585	50	51	840	0.96	0.94	0.94	0.97
train	GradientBoost	7472	136	404	3776	0.95	0.97	0.90	0.98
val	GradientBoost	1584	38	69	835	0.96	0.96	0.92	0.98
test	GradientBoost	1606	29	94	797	0.95	0.96	0.89	0.98
train	AdaBoostClf	7347	261	409	3771	0.94	0.94	0.90	0.97
val	AdaBoostClf	1561	61	77	827	0.95	0.93	0.91	0.96
test	AdaBoostClf	1571	64	93	798	0.94	0.93	0.90	0.96

Tabuľka 3.5: Vyhodnotenie predikcií všetkých modelov – delécie

Data	**Model**	**TN**	**FP**	**FN**	**TP**	**Presnosť**	**PPV**	**Senzitivita**	**Špecificita**
train	LogisticRegres	0	6417	0	2019	0.24	0.24	1.00	0.00
val	LogisticRegres	0	1385	0	423	0.23	0.23	1.00	0.00
test	LogisticRegres	0	1380	0	428	0.24	0.24	1.00	0.00
train	SVC	6354	63	521	1498	0.93	0.96	0.74	0.99
val	SVC	1368	17	113	310	0.93	0.95	0.73	0.99
test	SVC	1355	25	111	317	0.92	0.93	0.74	0.98
train	linearSVC	6416	1	2019	0	0.76	0.00	0.00	1.00
val	linearSVC	1384	1	423	0	0.77	0.00	0.00	1.00
test	linearSVC	1379	1	428	0	0.76	0.00	0.00	1.00
train	KneighborsClf	6298	119	320	1699	0.95	0.93	0.84	0.98
val	KneighborsClf	1351	34	69	354	0.94	0.91	0.84	0.98
test	KneighborsClf	1326	54	69	359	0.93	0.87	0.84	0.96
train	XGBoostClf	6406	11	53	1966	0.99	0.99	0.97	1.00
val	XGBoostClf	1363	22	40	383	0.97	0.95	0.91	0.98
test	XGBoostClf	1358	22	27	401	0.97	0.95	0.94	0.98
train	RandomForest	6416	1	0	2019	1.00	1.00	1.00	1.00
val	RandomForest	1369	16	37	386	0.97	0.96	0.91	0.99
test	RandomForest	1360	20	29	399	0.97	0.95	0.93	0.99
train	GradientBoost	6358	59	169	1850	0.97	0.97	0.92	0.99
val	GradientBoost	1372	13	43	380	0.97	0.97	0.90	0.99
test	GradientBoost	1362	18	33	395	0.97	0.96	0.92	0.99
train	AdaBoostClf	6306	111	196	1823	0.96	0.94	0.90	0.98
val	AdaBoostClf	1359	26	50	373	0.96	0.93	0.88	0.98
test	AdaBoostClf	1354	26	38	390	0.96	0.94	0.91	0.98

Tabuľka 3.6: Vyhodnotenie predikcií všetkých modelov – duplikácie

AttributeName	**Point Biserial Coefficient**	**PCB pozícia**	**feature_ importance**	**FI pozícia**
CGscore_indel_gene_count_2	0.54002	1	0.32102	1
gene_count	0.38597	2	0.03616	6
min_D_Frequency_3	-0.36703	3	0.05664	5
morbidGene_count	0.32697	4	0.24495	2
length	0.32283	5	0.19192	3
ExAC_zScore_pLI	0.29173	6	0.10269	4
ExAC_zScore_indel	0.23672	7	0.03073	7
HI_DDDperc_HI_gene_count	-0.03589	8	0.01588	8

Tabuľka 3.7: Zvolené atribúty pre delécie, Point-Biserial korelačný koeficient, pozície ku obrázku 3.3, hodnota podľa funkcie feature_importance

AttributeName	**Point Biserial Coefficient**	**PCB pozícia**	**feature_ importance**	**FI pozícia**
gene_count	0.50162	1	0.58852	1
length	0.45401	2	0.07000	4
morbidGene_count	0.44980	3	0.13012	3
ExAC_zScore_pLI	0.42318	4	0.13684	2
min_D_Frequency_3	-0.29399	5	0.03637	6
ExAC_zScore_indel	0.28275	6	0.03814	5

Tabuľka 3.8: Zvolené atribúty pre duplikácie, Point-Biserial korelačný koeficient, pozície ku obrázku 3.4, hodnota podľa funkcie feature_importance

Príloha B: obsah elektronickej prílohy

V elektronickej prílohe priloženej k práci sa nachádzajú nasledovné: (zverejnené aj na `https://github.com/MiskaGaziova/bakalarka_isv`

- data_annotsv:
  - AnnotSV.r2yo0QbkMl_full.tsv súbor s CNV anotáciami (z nástroja AnnotSV)
  - train_val_test priečinok obsahuje trénovacie, validačné a testovacie dáta

Skripty/Jupyter Notebooky:

- attribute_correlations.ipynb – porovnávanie atribútov
- data_prepare.ipynb, ingest_data.ipynb – vytvorenie dátových sád a nových atribútov
- train_evaluate_models.ipynb, model_evaluation.ipynb – vyhodnotenie presností modelov
- unclassified_regions.ipynb – hľadanie oblastí neklasifikovaných CNV
- porovnanie_ClassifyCNV_AnnotSV.ipynb – vyhodnotenie presností ClassifyCNV a AnnotSV na testovacej sade

Priečinky s výsledkami:

- attributes_selection – tabuľky a grafy vybraných atribútov ako korelujú s klinickým významom (podľa Point-Biserial korelačného koeficientu), ako sa správa prenosť modelu postupným pridávaním vybratých atribútov
- comparison_attributes_annotsv_full – porovnania atribútov ohodnotením CNV ako celku
  - correlations_attributes – tabuľky a grafy znázorňujúce korelácie atribútov podľa Pearsonovho a Point-Biserial korelačného koeficientu
  - *correlations_info_train_loss_gain.tsv* – komplexné hodnotenie atribútov pre delécie aj duplikácie (p-hodnota podľa Mann-Whitney Testu, Point-Biserial koeficient, min, max, medián, priemer)

- comparison_attributes_annotsv_split – porovnania atribútov génových hodnôt
- models – natrénované modely, tabuľky a vizualizácie presností viacerých modelov na jednotlivých dátových sadách
- test_prediction – tabuľky a vizualizácie predikcií samotného modelu aj klasikifikácií neklasifikovaných CNV
- unclassified_tresholds - grafy vygenerované pri identifikácii neklasifikovaných CNV
- classifyCNV_test – klasifikácie CNV z testovacích sád nástrojom ClassifyCNV