



Model dubokog učenja za određivanje očitanih baza dobivenih uređajem za sekvenciranje MinION

Autor: Marko Ratković Mentor: izv. prof. dr. sc. Mile Šikić
Sveučilište u Zagrebu,
Fakultet elektrotehnike i računarstva
Zavod za elektroničke sustave i obradu informacija

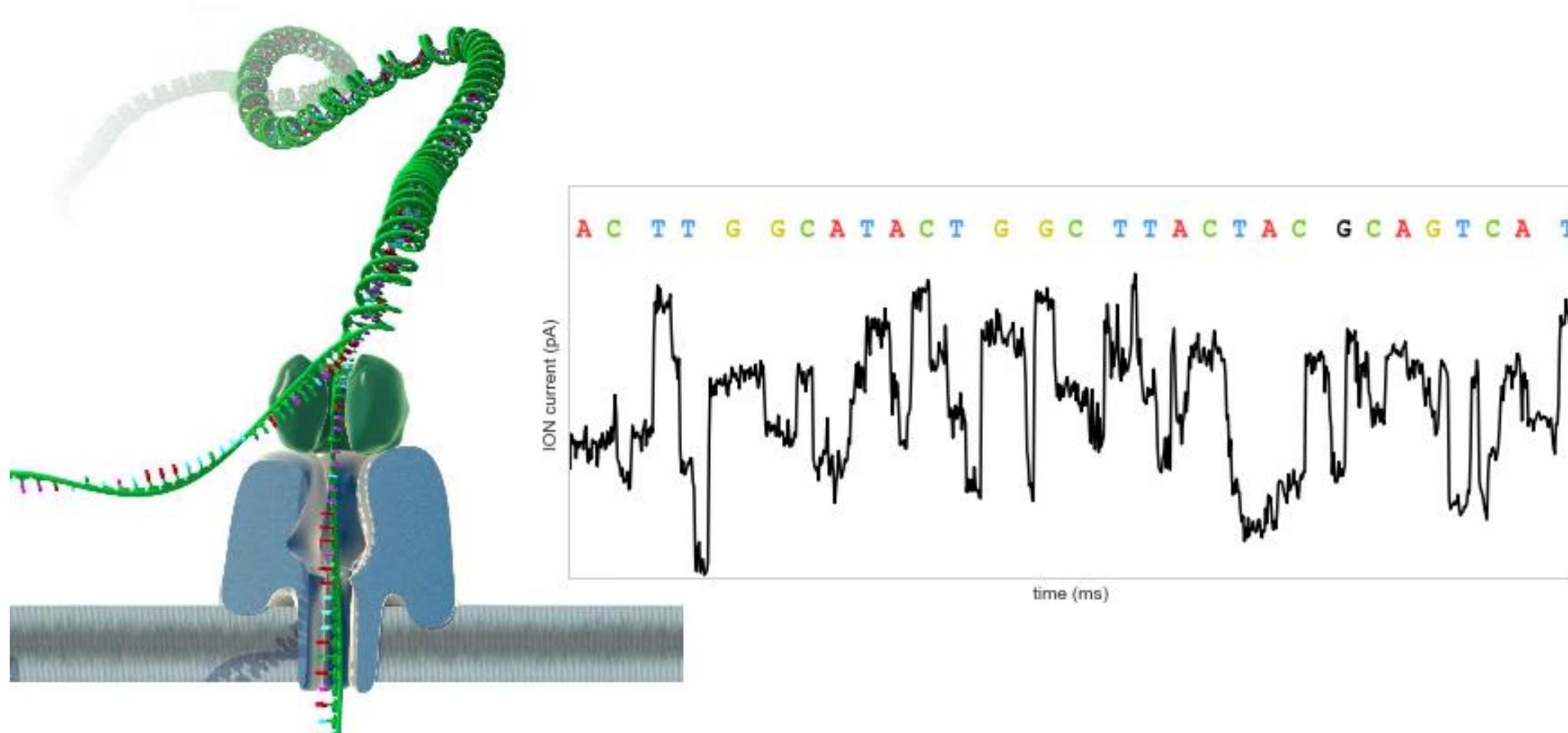


1. Uvod

Oxford Nanopore Technologies (ONT) 2015. godine predstavio je novi uređaj za sekvenciranje. **MinION** je veličine dlana, teži svega 90 grama, poveziv putem USB 3.0, cijene \$1000 dolara, generira dugačka očitavanja, no uz vrlo veliku pogrešku. Cilj ovog rada je pokazati da pogreška ne ovisi samo o tehnologiji sekvenciranja već i programskoj podršci koja analizira izlaz samog instrumenta te ga pretvara u niz nukleotida te je istu moguće poboljšati korištenjem metoda strojnog učenja.

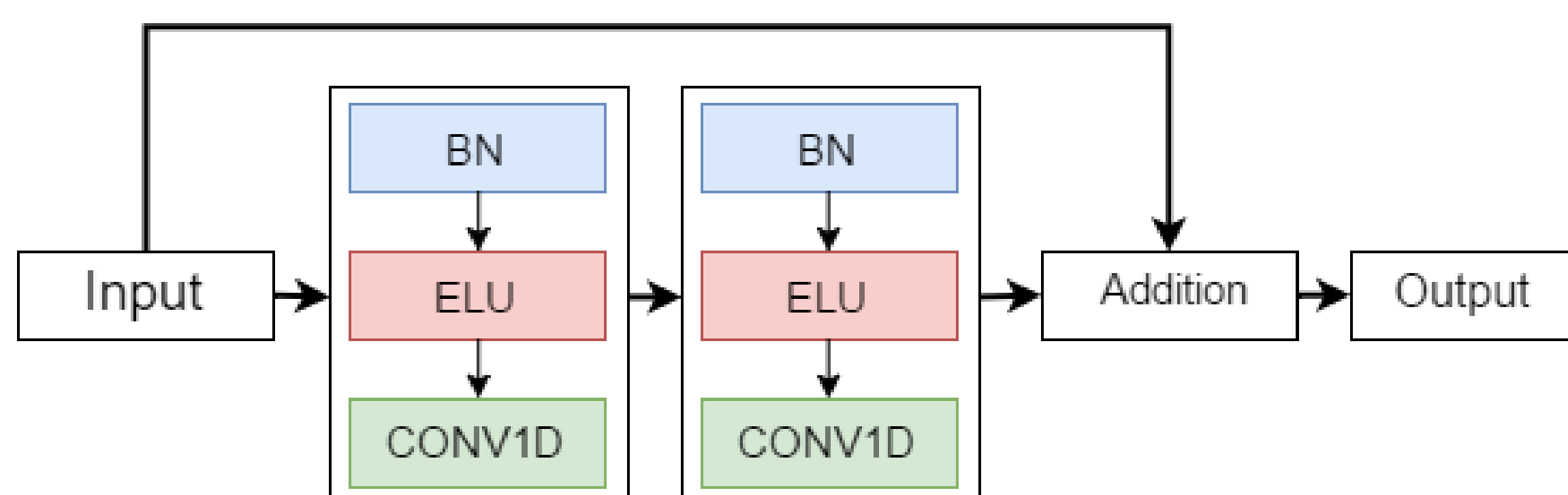
2. Opis problema

Prolaskom DNA kroz nanoporu, mijenja se električni otpor ovisno o nukleotidima koji se nalaze unutar same pore. Dobiveni signal, odnosno uzorci struje, opisuju sekvencirani DNA lanac.



3. Metoda

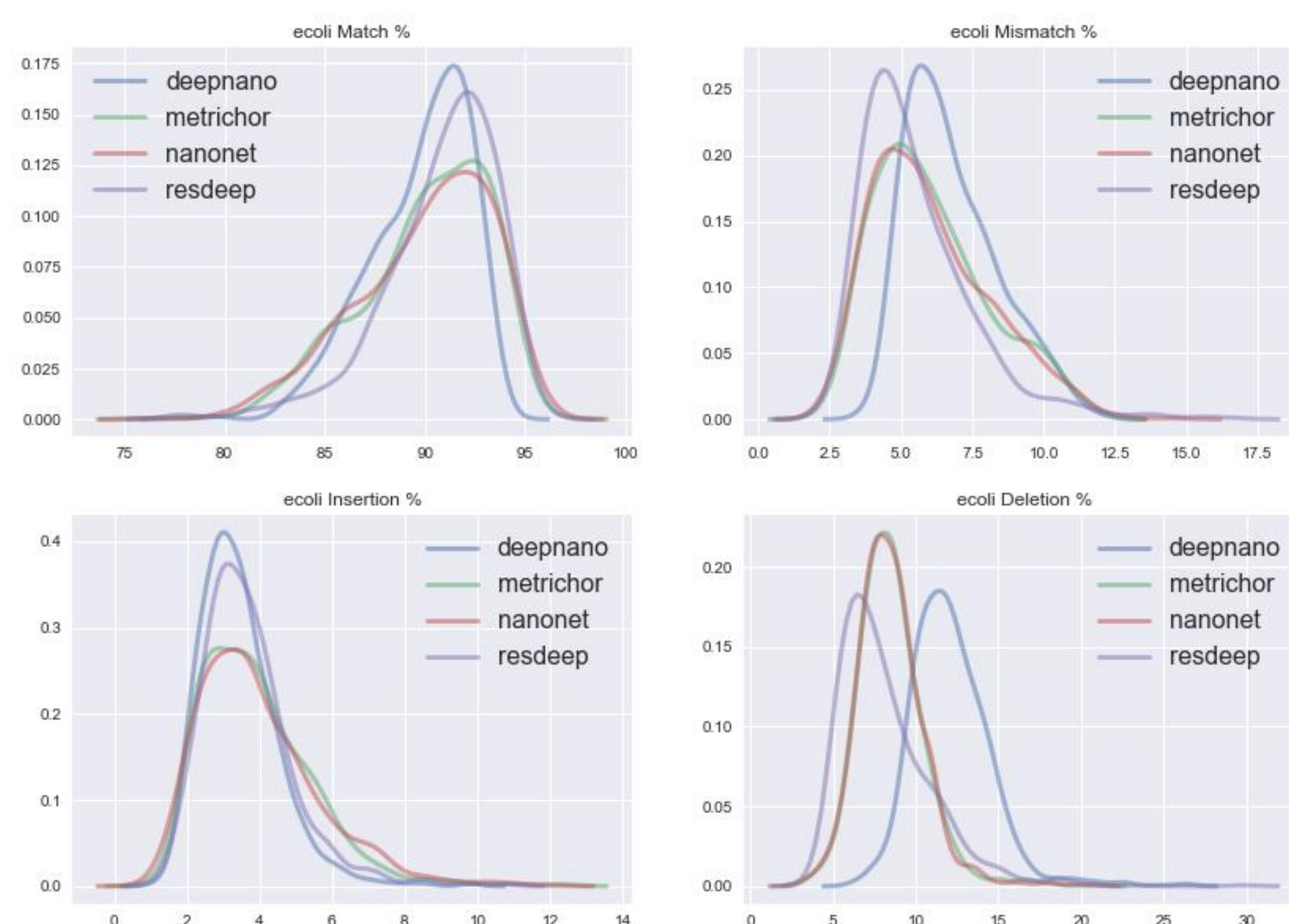
Opisani problem se može shvatiti kao **problem strojnog prevođenja** gdje se mjerenja struje prevode u nukleotide, slova A, C, T i G. Konačni model je **rezidualna konvolucijska neuronska mreža** koja se sastoji od 72 bloka sa slike. Rezidualna arhitektura, **ELU** aktivacija i **batchnorm** otklanjaju problem iščezavajućeg gradijenta.



Connectionist Temporal Classification (CTC) gubitak je korišten kako bi se riješio problem varijabilne duljine izlaznih sekvenci i potrebe za poravnanjem željenih sekvenci na izlaze prilikom treniranja. Model je treniran nad očitanjima dobivenim službenim alatom dodatno *popravljenim* poravnanjem na referentni genom.

4. Rezultati

Kvaliteta očitavanja evaluirana je na nekoliko načina: poravnavanjem na referencu, određivanjem konsenzusa *de novo* sastavljanjem ili iz *pileupa* poravnanja i usporedbom tako dobivenih konsenzusa s referencom. Razvijeni model, u nastavku pod imenom **resdeep** nazvan po korištenoj arhitekturi, pokazao je bolje rezultate od postojećih alata (službeni **Metrichor** i **Nanonet** te *third-party* **DeepNano**) u svim aspektima, na oba korištena skupa podataka (*E. Coli* i *lambda*)



KDE plot operacija poravnanja za očitavanja

	Poklapanja	Zamjene	Umetanja	Brisanja
<i>resdeep</i>	99.24	0.65	0.11	0.55
<i>Nanonet</i>	97.97	1.57	0.46	1.52
<i>DeepNano</i>	98.87	1.00	0.12	0.90
<i>Metrichor</i>	99.12	0.75	0.13	0.63

Postotak operacija u konsenzus sekvenci

	CPU	GPU
<i>resdeep</i>	1363.4	6571.7
<i>Nanonet</i>	897.5	3828.4
<i>DeepNano</i>	692.7	/
<i>Metrichor</i>	online servis	

Brzina izražena u bazama po sekundi [bp/s]

5. Zaključak

Predstavljeno je rješenje za očitavanje baza temeljeno na konvolucijskim neuronskim mrežama koje pruža napredak u preciznosti i brzini u odnosu na postojeća rješenja. Razvijeni model trenutno podržava podatke nastale R9 kemijom ali je moguće proširenje na R9.4 te R9.5 kad podaci postanu dostupni.