Heaven’s Light is Our Guide

DEPRATMENT OF COMPUTER SCIENCE AND ENGINEERING

Rajshahi University of Engineering & Technology

Bioactivity Prediction from Target Proteins using Machine Learning Models

Presented by	Supervised by
Tasneem Sakif Ibne Alam	**Prof. Dr. Md. Rabiul Islam**
Roll: 1603094	**Department of CSE, RUET**
Department of CSE, RUET	

OUTLINES

INTRODUCTION

Target Protein

- Root cause of a disease.
- Spreads the disease creating protein pathways.

Drug

- Interacts with target proteins.
- Can control the course of a disease.

Drug Discovery

- To find out drug hits.
- Drug hits: Compounds which are expected to alter the disease.

OBJECTIVES

- To work with two datasets of different sizes.
- To preprocess the datasets to make those more uniform.
- To extract features or molecular descriptors from the compounds.
- To apply machine learning models to the datasets.
- To analyze and compare the results of the models.
- To find out the best model.

LITERATURE REVIEW

1. Validating the validation: reanalyzing a large-scale comparison of deep learning and machine learning models for bioactivity prediction [1] (Matthew C. Robinson, Et al.)

Contribution

- Worked with datasets of different sizes.
- Showed SVM and FNN are the best models

Limitations

- Dataset is not uniform as didn’t differentiate between IC50, EC50, potency etc. values.
- Problems with classification of actives/inactives.

LITERATURE REVIEW (CONT.)

Table 1: Matthew C. Robinson Et al.'s experimental results using 3 Fold cross validation [1]

		Fold 1	Fold	Fold	Mean	SEM
A: ChEMBL 1964055	FNN AUC–ROC (95% CI)	0.44 (0.035, 0.94)	0.62 (0.0, 1.0)	0.64 (0.34, 0.86)	0.57	0.05
	SVM AUC–ROC (95% CI)	0.38 (0.02, 0.94)	0.97 (0.0, 1.0)	0.68 (0.38, 0.88)	0.67	0.14
	Test set size (actives/ inactives)	35 (32/3)	30 (29/1)	35 (29/6)		
B: ChEMBL 1794580	FNN AUC–ROC (95% CI)	0.889 (0.883, 0.895)	0.905 (0.900, 0.910)	0.906 (0.900, 0.911)	0.900	0.005
	SVM AUC–ROC (95% CI)	0.936 (0.921, 0.931)	0.926 (0.921, 0.930)	0.934 (0.930, 0.939)	0.929	0.002
	Test set size (actives/ inactives)	19388 (5553/13885)	25165 (6918/18247)	19363 (5491/13872)		

LITERATURE REVIEW (CONT.)

2. Large-scale comparison of machine learning methods for drug target prediction on ChEMBL [2] (Andreas Mayr, Et al.)

Contribution

- Worked with large scale datasets.
- Deep learning model gave the best accuracy.

Limitations

- Data was not preprocessed properly.
- Imbalanced data.
- Same model performed differently for different datasets.

LITERATURE REVIEW (CONT.)

Table 2: Andreas Mayr Et al.'s experimental results (partial) using 2 Fold cross validation [2]

Assay	Surrgate Assay	Target	Surrogate Assay Accuracy	Deep learning accuracy
CHEMBL1909134	CHEMBL1613777	CYP450-2C19	0.54 [0.4136, 0.653]	0.95 [0.9198, 0.9658]
CHEMBL1909200	CHEMBL1614521	ERK	0.56 [0.4012, 0.7005]	0.98 [0.9615, 0.9912]
CHEMBL1963940	CHEMBL1794352	Luciferase	1.00 [0.8076, 1]	0.87 [0.775, 0.9344]
CHEMBL1741321	CHEMBL1614110	CYP450-2D6	0.99 [0.9889, 0.9956]	0.83 [0.8184, 0.8352]
CHEMBL1741325	CHEMBL1614027	CYP450-2C9	0.99 [0.9839, 0.993]	0.75 [0.7428, 0.762]
CHEMBL1741323	CHEMBL1614027	CYP450-2C19	0.99 [0.9822, 0.9911]	0.77 [0.7602, 0.7789]

PROPOSED METHODOLOGY

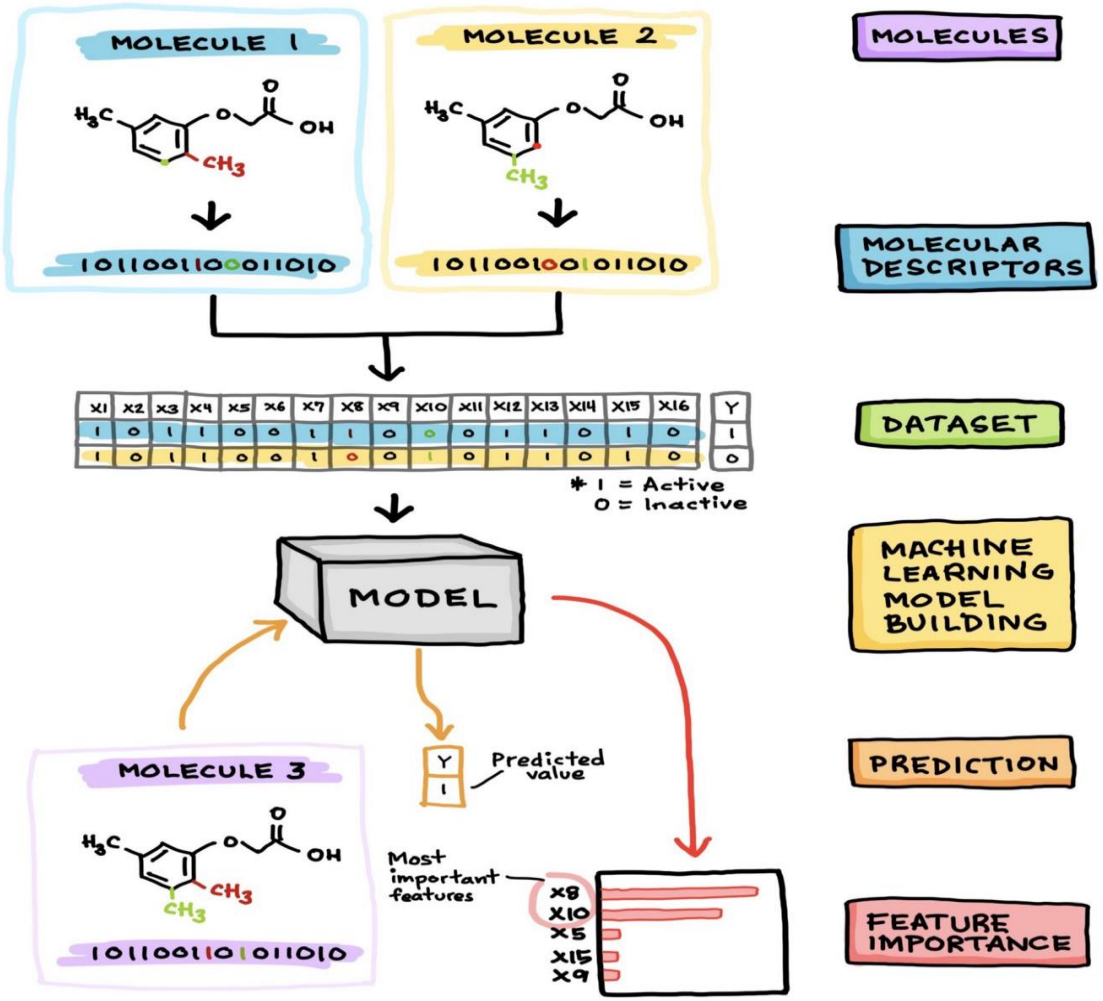


Figure 1: Computational drug discovery overview [3].

PROPOSED METHODOLOGY (CONT.)

Datasets

- Collected from Chembl database [4].
- Plasmodium Falciparum (Chembl id: 364): 43324 compounds.
- Leukocyte Elastase (Chembl id: 248): 3147 compounds.

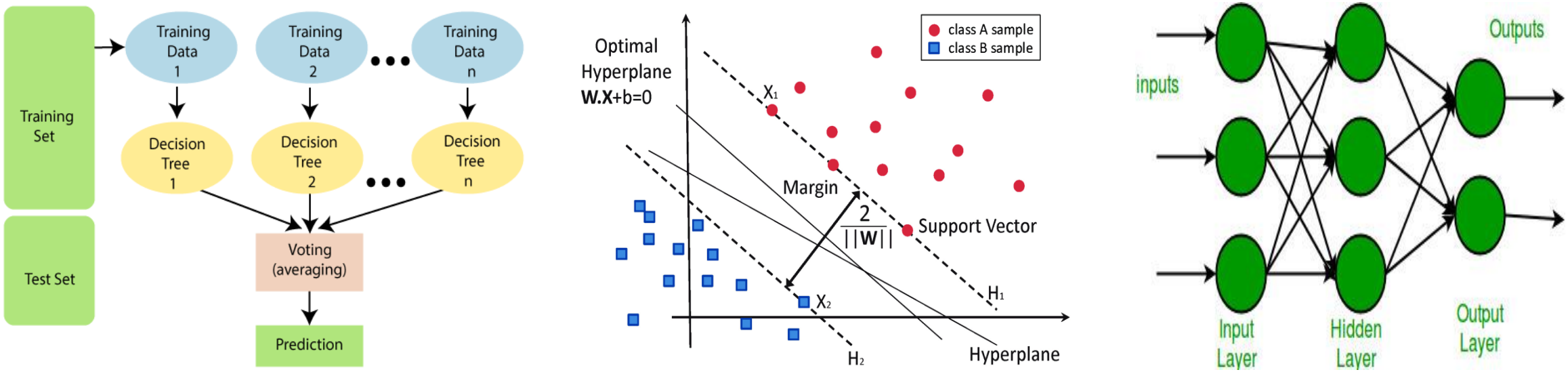
Feature Extraction

- Using PaDEL descriptor [5].
- In the form of PubChem fingerprints.

Machine learning models

- Random Forest (RF) Classifier.
- Support Vector Machine (SVM) Classifier.
- Feed Forward Neural Network (FNN) model.

PROPOSED METHODOLOGY (CONT.)

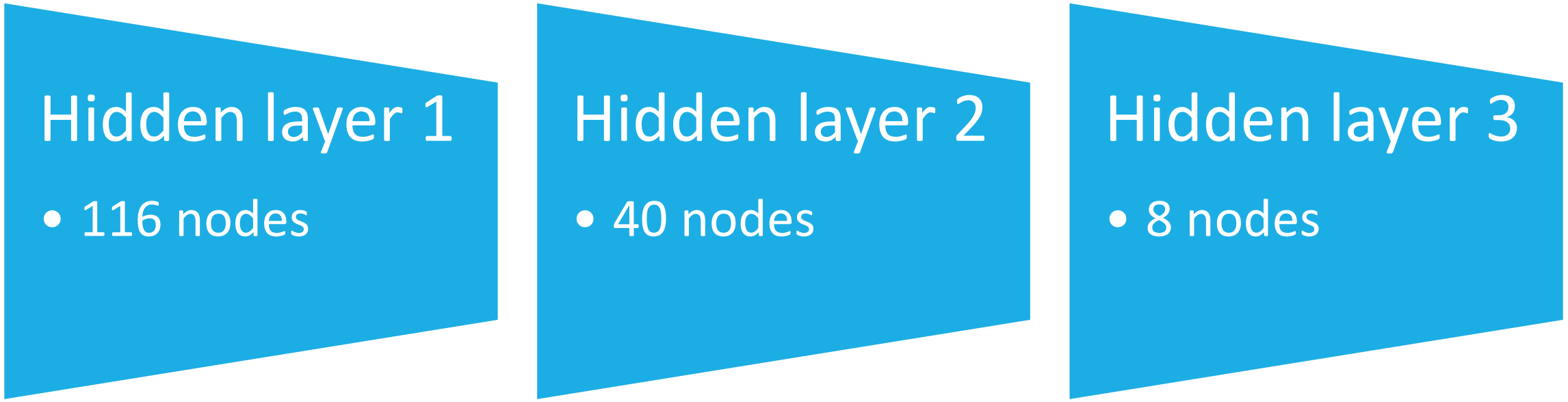


(a) (b) (c)

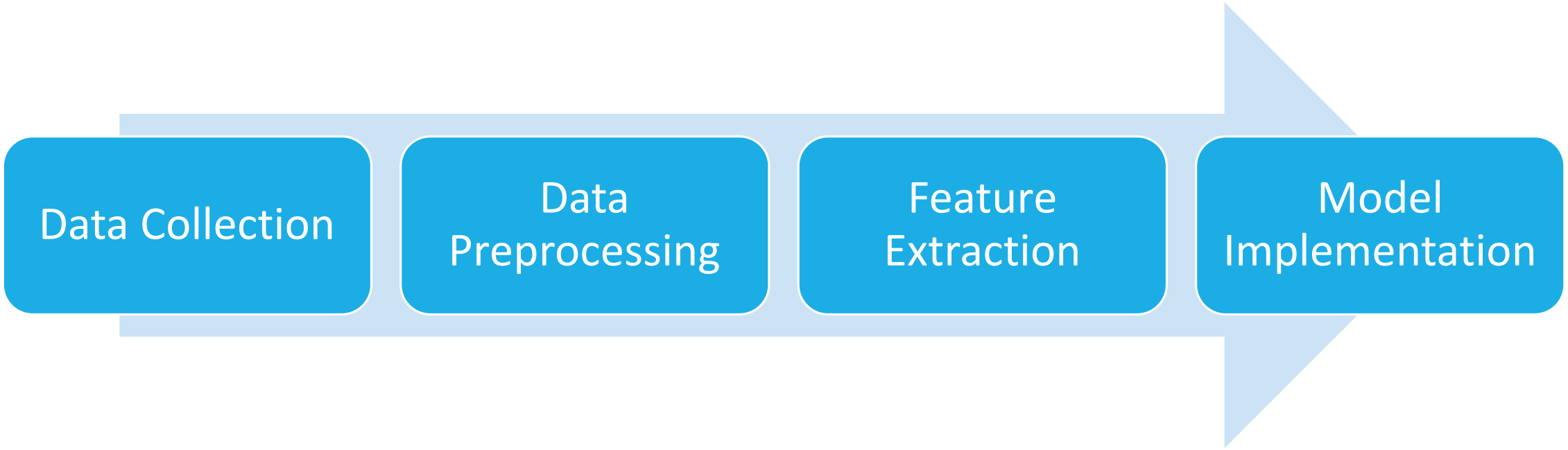
Figure 2: (a) Random Forest [6] (b) Support Vector Machine [7] (c) Feed Forward Neural Network [8].

PROPOSED METHODOLOGY (CONT.)

Proposed Feed Forward Neural Network architecture:



IMPLEMENTATION



IMPLEMENTATION (CONT.)

Data Preprocessing

- Handling the missing data.
- Rows having missing columns of standard value and canonical smiles were dropped.
- Removing duplicate canonical smiles values.
- Converting IC50 to PIC50.
- Actives and Inactives were classified.
- Active: Standard value ≤ 1000 nM
- Inactive: Standard value > 1000 nM

IMPLEMENTATION (CONT.)

Table 3: First few entries of the preprocessed dataset of Plasmodium Falciparum

Id	molecule_chembl_id	canonical_smiles	class	pIC50
0	CHEMBL77052	C[C@@H]1CC[C@H]2[C@@H](C)[C…	active	8.37161
1	CHEMBL307145	Oc1cccc(O)c1O	inactive	5.24718
2	CHEMBL16300	O=C(NO)c1ccccc1	inactive	4.75448
3	CHEMBL307153	C[C@@H]1CC[C@H]2[C@@H](C….	active	8.01999
4	CHEMBL339049	CC(C)(C)NCc1cc(Nc2ccnc3cc(Cl)….	active	7.74472
5	CHEMBL316098	CC(C)(C)NCc1cc(Nc2ccnc3cc(Cl)….	active	8.83268
6	CHEMBL93286	CC(C)(C)NCc1cc(Nc2ccnc3cc(….	active	8.16749
7	CHEMBL337981	CC(C)(C)NCc1cc(Nc2ccnc3cc(Cl)..	active	7.23657

IMPLEMENTATION (CONT.)

Feature Extraction

- Calculated PubChem fingerprint descriptors using PaDEL descriptor software.
- PubChem fingerprint encodes molecular fragments information with 881 binary digits [9].
- The 881 binary digits indicates absence or presence of certain features in the molecules.
- PubChem is useful for similarity neighboring and similarity searching.
- Removing the low variance features we worked with 174 features.

IMPLEMENTATION (CONT.)

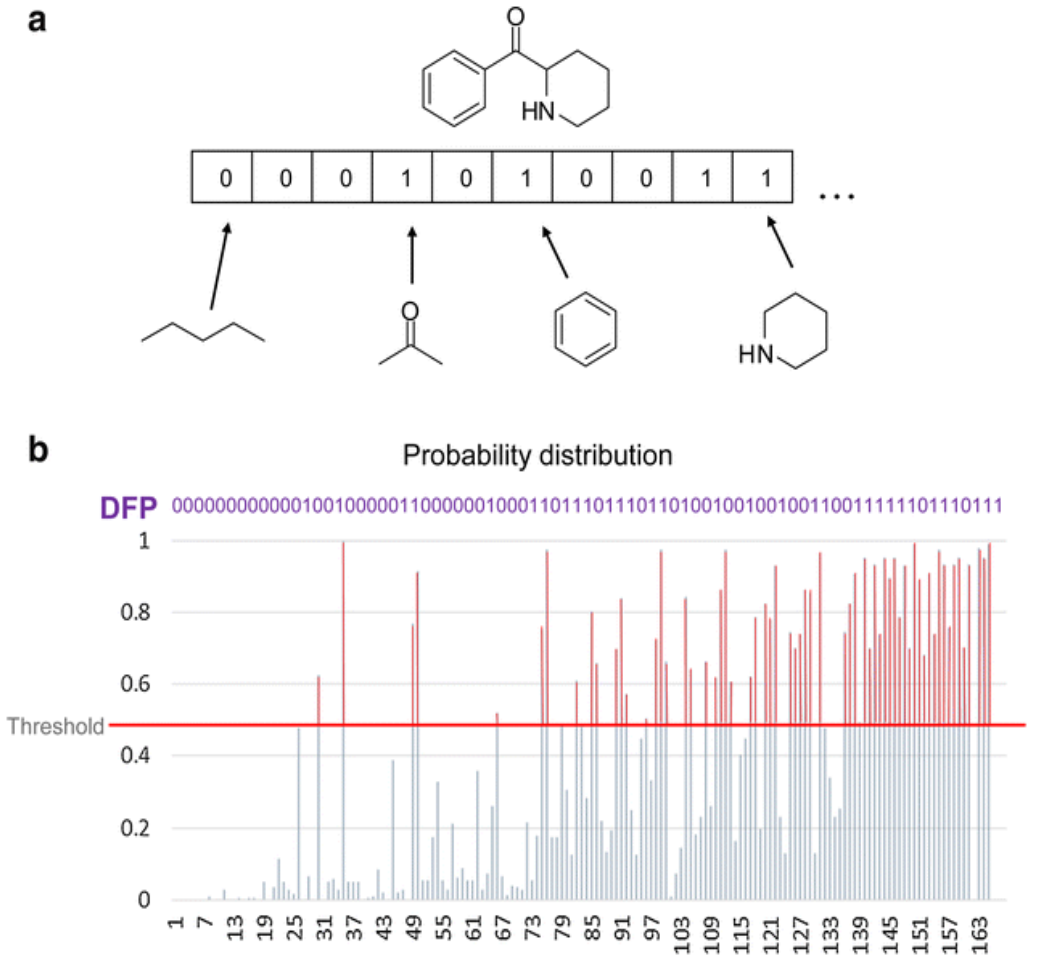


Figure 3: (a) Schematic representation of a binary and dictionary-based molecular fingerprint. (b) Schematic representation of a database fingerprint (DFP) [9].

IMPLEMENTATION (CONT.)

Model Implementation

RF and SVM

- Split the data 80/20 for testing and training.
- Fed the data into the classification models.
- Found out accuracies.
- Generated Confusion matrices.
- Printed out the classification reports.

FNN

- Split the data 80/20 for testing and training.
- Set ReLU activation function for all the hidden layers.
- Set sigmoid activation function for output layer.
- Loss: Binary Cross Entropy
- Optimizer: Adam
- Fed the data.
- Batch size:32, Epoch: 200

EXPERIMENTAL RESULTS AND PERFORMANCE ANALYSIS

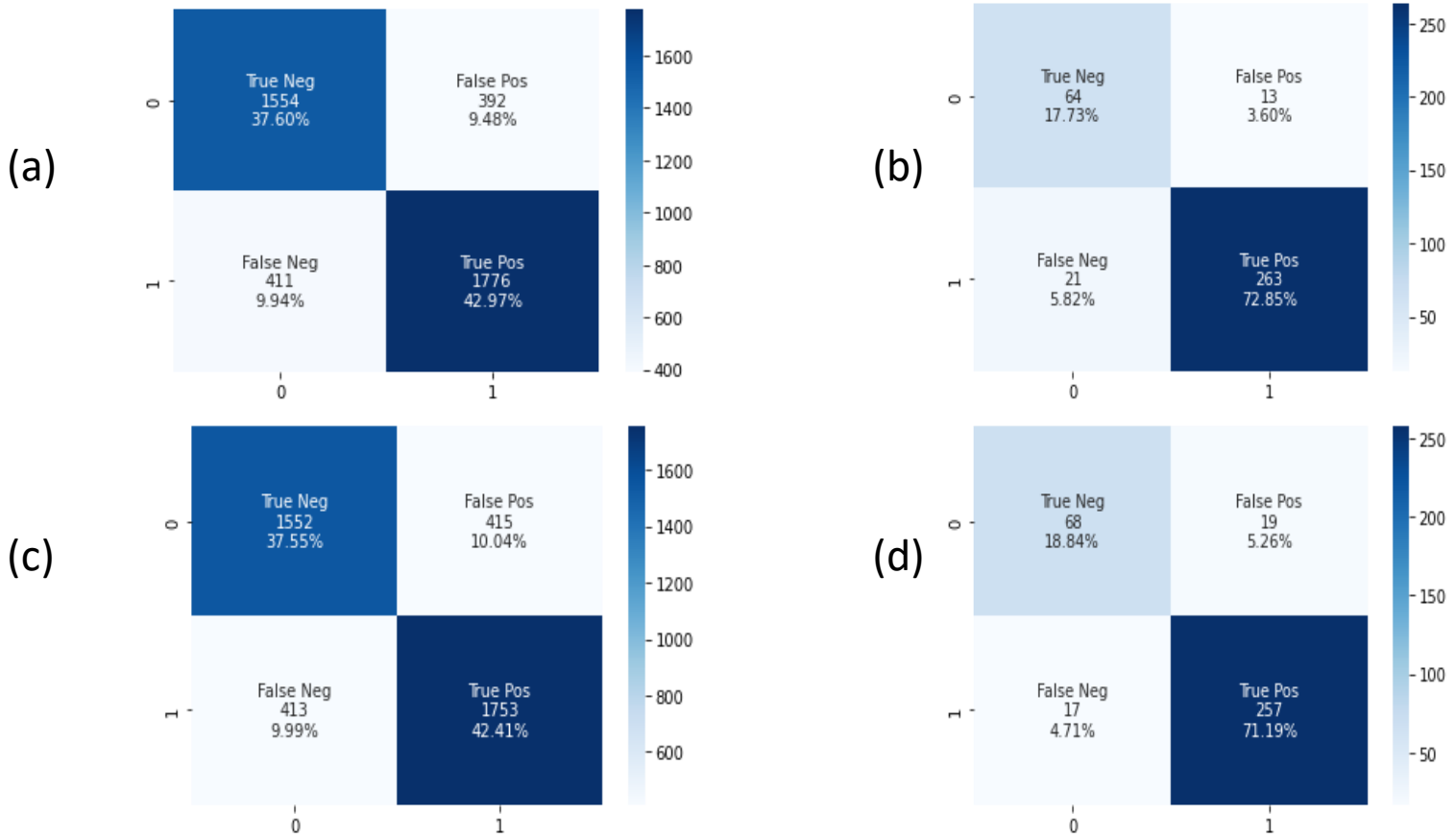


Figure 4: Confusion matrices of (a) RF on Plasmodium Falciparum (b) RF on Leukocyte Elastase (c) SVM on Plasmodium Falciparum (d) SVM on Leukocyte Elastase dataset

EXPERIMENTAL RESULTS AND PERFORMANCE ANALYSIS (CONT.)

Table 4: Evaluation metrics of (a) RF on Plasmodium Falciparum (b) RF on Leukocyte Elastase (c) SVM on Plasmodium Falciparum (d) SVM on Leukocyte Elastase dataset

(a)

	Precision	Recall	F1-score	Support
0	0.79	0.80	0.79	1946
1	0.82	0.81	0.82	2187
Accuracy	-	-	0.81	4133
Macro avg.	0.81	0.81	0.81	4133
Weighted avg.	0.81	0.81	0.81	4133

(b)

	Precision	Recall	F1-score	Support
0	0.80	0.78	0.79	87
1	0.93	0.94	0.93	274
Accuracy	-	-	0.91	361
Macro avg.	0.87	0.86	0.86	361
Weighted avg.	0.90	0.90	0.90	361

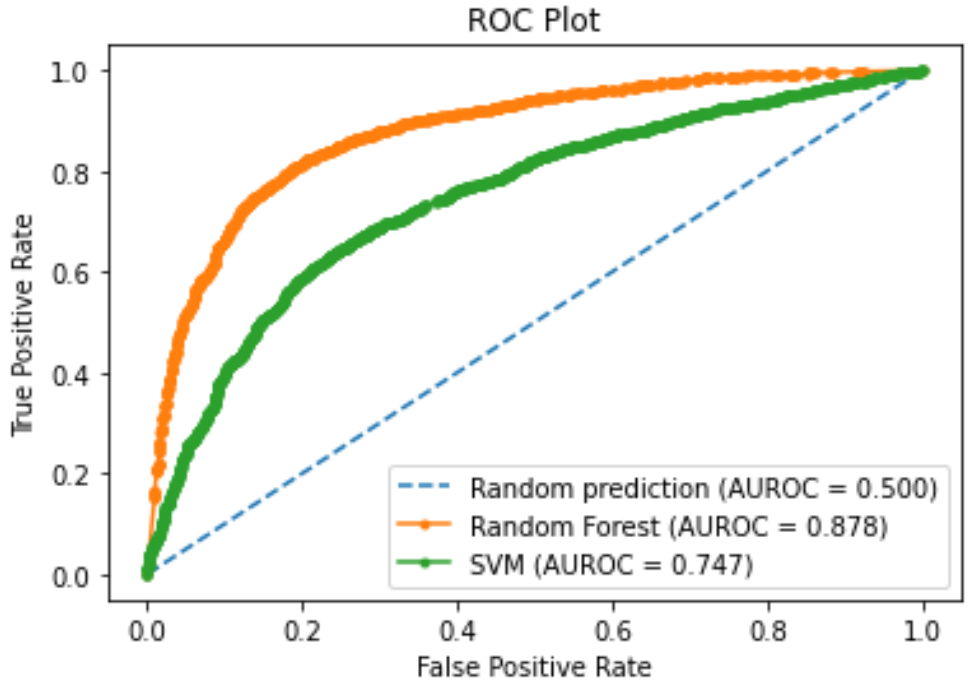
(c)

	Precision	Recall	F1-score	Support
0	0.79	0.79	0.79	1967
1	0.81	0.81	0.81	2166
Accuracy	-	-	0.80	4133
Macro avg.	0.80	0.80	0.80	4133
Weighted avg.	0.80	0.80	0.80	4133

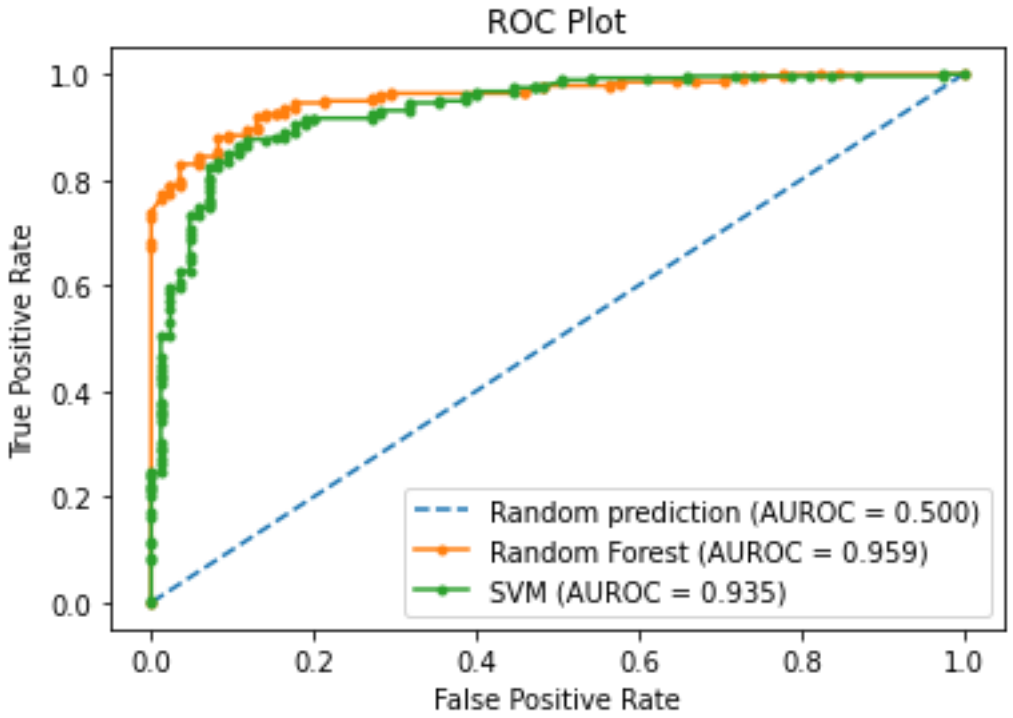
(d)

	Precision	Recall	F1-score	Support
0	0.82	0.80	0.81	88
1	0.93	0.95	0.94	273
Accuracy	-	-	0.91	361
Macro avg.	0.88	0.87	0.88	361
Weighted avg.	0.91	0.91	0.91	361

EXPERIMENTAL RESULTS AND PERFORMANCE ANALYSIS (CONT.)



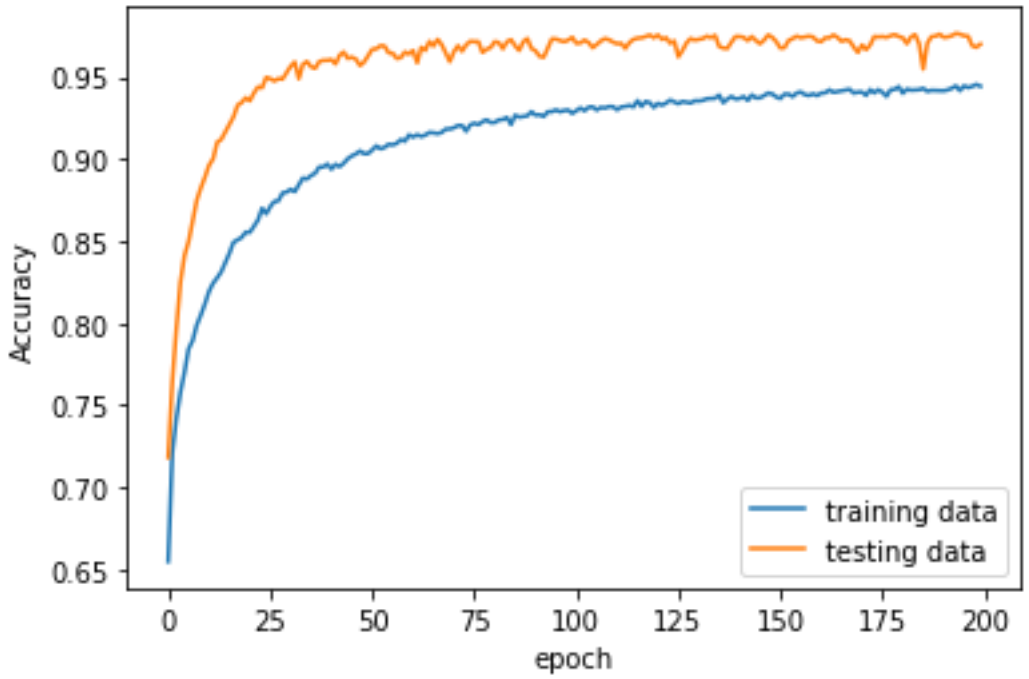
(a)



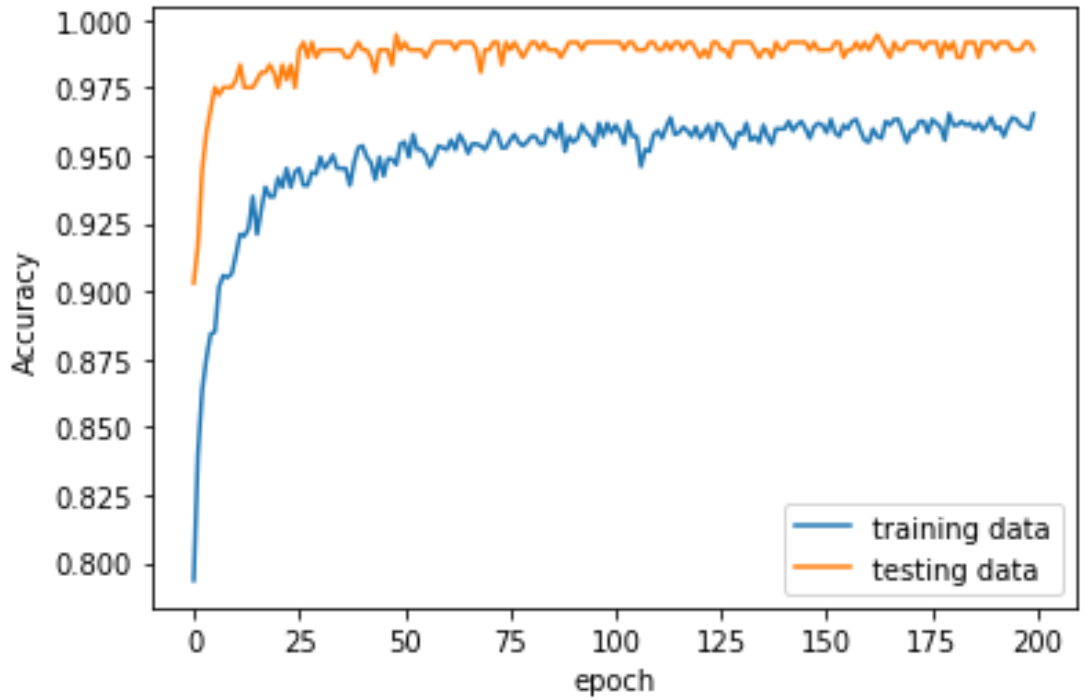
(b)

Figure 5: RF versus SVM ROC curve for (a) Plasmodium Falciparum (b) Leukocyte Elastase dataset

EXPERIMENTAL RESULTS AND PERFORMANCE ANALYSIS (CONT.)



(a)



(b)

Figure 6: FNN's Epoch versus Accuracy curve for (a) Plasmodium Falciparum (b) Leukocyte Elastase dataset

EXPERIMENTAL RESULTS AND PERFORMANCE ANALYSIS (CONT.)

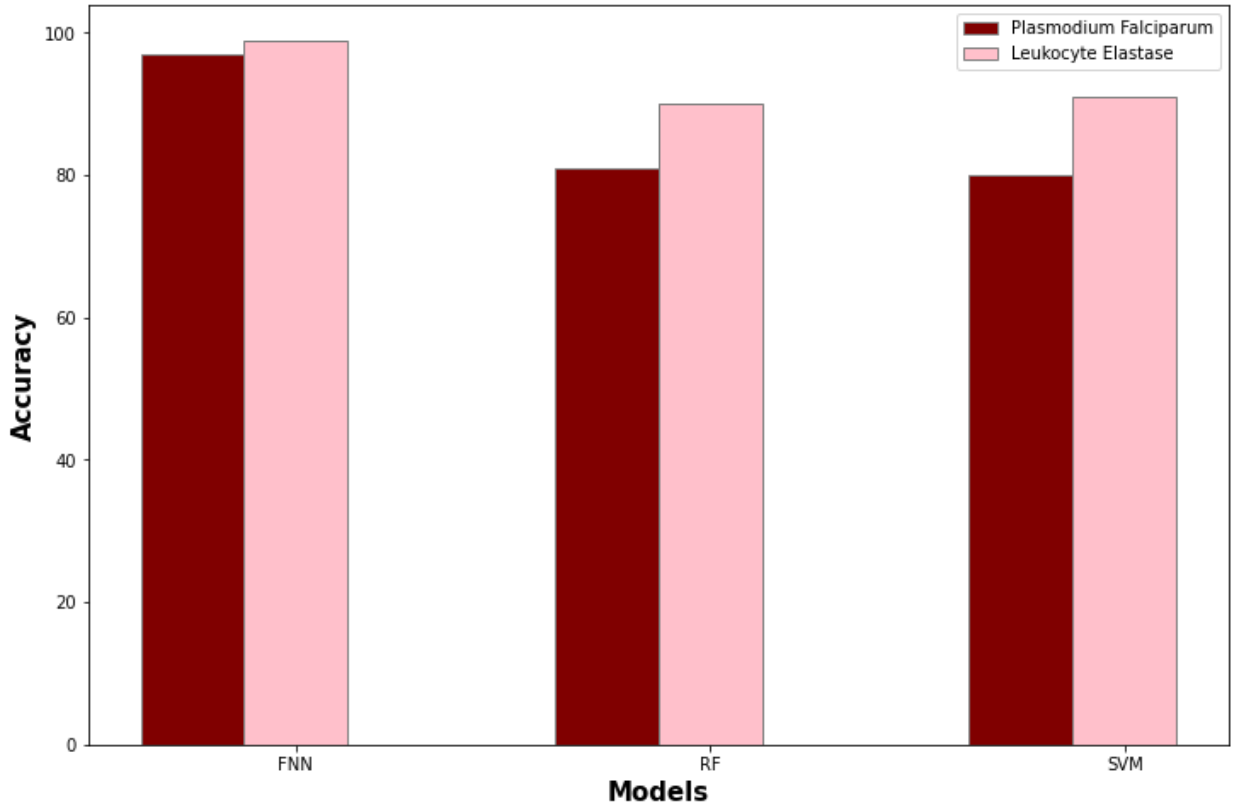


Figure 7: Model Comparison for both Dataset

Observation & Analysis

- FNN outperforms both RF and SVM for both datasets.
- Performance of RF and SVM significantly downgrades for larger data samples.
- FNN keeps up its good performance even for larger data samples.
- FNN is the best performing model.

EXPERIMENTAL RESULTS AND PERFORMANCE ANALYSIS (CONT.)

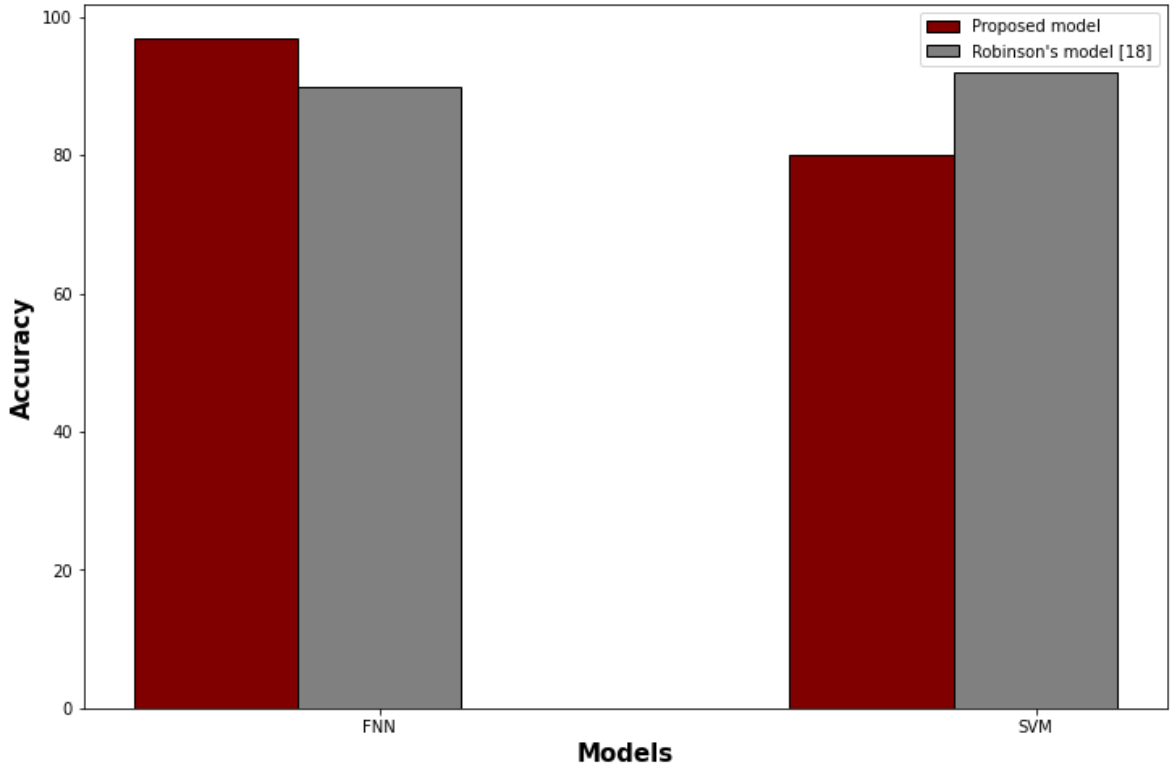


Figure 8: Proposed model's result comparison with Matthew C. Robinson Et al.'s models [1]

Observation & Analysis

- Proposed FNN outperforms Robinson's [1] FNN model with 97% vs. 92% accuracy scores.
- Robinson's [1] SVM with 92% accuracy outperforms our SVM with 80% accuracy.
- Overall our proposed FNN model gives the best accuracy.
- Difference in results because our dataset is more balanced and uniform.

CONCLUSION AND FUTURE WORKS

Conclusion

- FNN turned out to be the best performing model.
- SVM and RF lacks performance for large dataset.

Future Works

- To work with different datasets.
- To work with even larger datasets.
- To use different molecular descriptors as features.
- To implement different machine learning and deep learning models.

REFERENCES

1. Robinson, M., Glen, R. and Lee, A., 2020. Validating the validation: reanalyzing a large-scale comparison of deep learning and machine learning models for bioactivity prediction. *Journal of Computer-Aided Molecular Design*, 34(7), pp.717-730.
2. Mayr, A., Klambauer, G., Unterthiner, T. and Steijaert, M., 2018. Large-scale comparison of machine learning methods for drug target prediction on ChEMBL. *Chemical Science*, 9(24), pp.5441-5451.
3. C. Nantasenamat, "How to build your first machine learning model using no code," *Medium*, 25-Jun-2021. [Online]. Available: https://towardsdatascience.com/how-to-build-your-first-machine-learning-model-using-no-code-a7cf8db37dd1. [Accessed: 10-Oct-2022].
4. "Chembl Database," *EBI*. [Online]. Available: https://www.ebi.ac.uk/chembl/. [Accessed: 10-Oct-2022].
5. Yap, C., 2010. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *Journal of Computational Chemistry*, 32(7), pp.1466-1474.

REFERENCES (CONT.)

6. “Machine learning random forest algorithm - javatpoint,” *www.javatpoint.com*. [Online]. Available: https://www.javatpoint.com/machine-learning-random-forest-algorithm. [Accessed: 10-Oct-2022].
7. García-Gonzalo, E., Fernández-Muñiz, Z., García Nieto, P., Bernardo Sánchez, A. and Menéndez Fernández, M., 2016. Hard-Rock Stability Analysis for Span Design in Entry-Type Excavations with Learning Classifiers. *Materials*, 9(7), p.531.
8. “Multi-layer perceptron learning in tensorflow,” *GeeksforGeeks*, 05-Nov-2021. [Online]. Available: https://www.geeksforgeeks.org/multi-layer-perceptron-learning-in-tensorflow/. [Accessed: 18-Oct-2022].
9. Fernández-de Gortari, E., García-Jacas, C., Martinez-Mayorga, K. and Medina-Franco, J., 2017. Database fingerprint (DFP): an approach to represent molecular databases. *Journal of Cheminformatics*, 9(1).