Prediction of Antimicrobial Peptides Using Machine Learning Approach

S.A.S Nagar

Mushtaq Ahmad Wani and Prabha Garg*

Department of Pharmacoinformatics, National Institute of Pharmaceutical Education and Research, S.A.S. Nagar, Punjab 160062, India

Introduction

Antimicrobial peptides (AMPs), also called host defense peptides (HDPs), consist of 12 to 100 amino acids that are part of the innate immune system and can be found among all classes of life including bacteria, fungi, plants, invertebrates, and vertebrates. These AMPs have been found to be effective against disease-causing pathogens. Identification of antimicrobial peptides through *in vitro* and *in vivo* experiments on large number of peptides is an expensive and time-consuming approach.

This study explores machine learning classifiers for predicting antimicrobial peptides (AMPs) using a diverse set of AMPs (2638) and non-AMPs (3700). The RF classifier-based model outperformed other models in both internal and external validations. It correctly predicted known AMPs and non-AMPs, with ChargeD2001, PAAC12 (pseudo amino acid composition), and polarity T13 being crucial features in AMPs' antimicrobial activity. The developed RF-based classification model may be useful in designing and predicting novel potential AMPs.

Aim & Objectives

Aim: Development of a predictive model for antimicrobial peptides using machine learning approach.

Objectives:

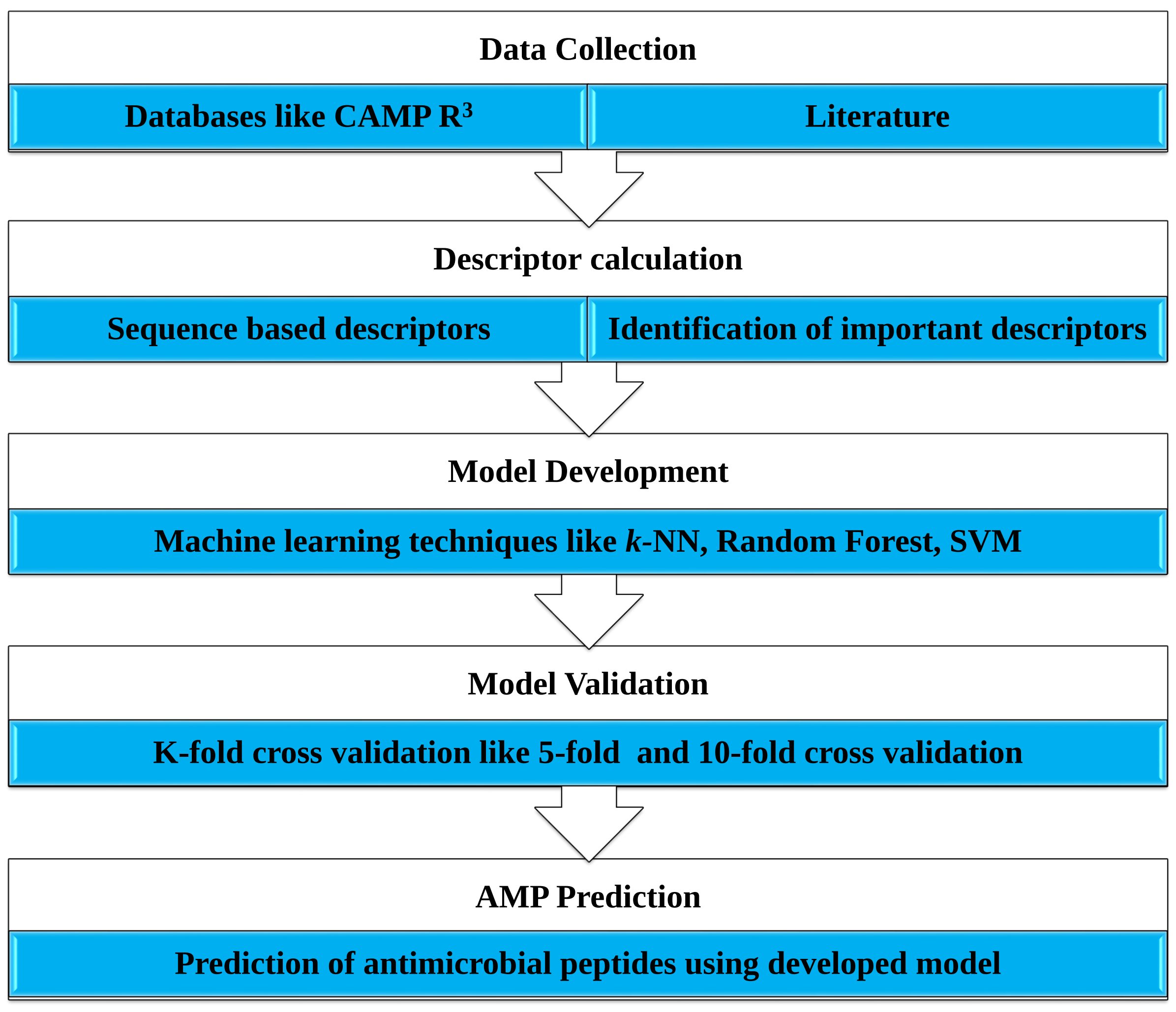
1.Collection of antimicrobial peptide data from various databases

2. Calculation of sequence based descriptors for the antimicrobial peptides

3. Identification of important descriptors for model development

4. Development and validation of models for predicting antimicrobial peptides

Workflow



Data Collection

Databases like CAMP R^3 | Literature

Descriptor calculation

Sequence based descriptors | Identification of important descriptors

Model Development

Machine learning techniques like *k*-NN, Random Forest, SVM

Model Validation

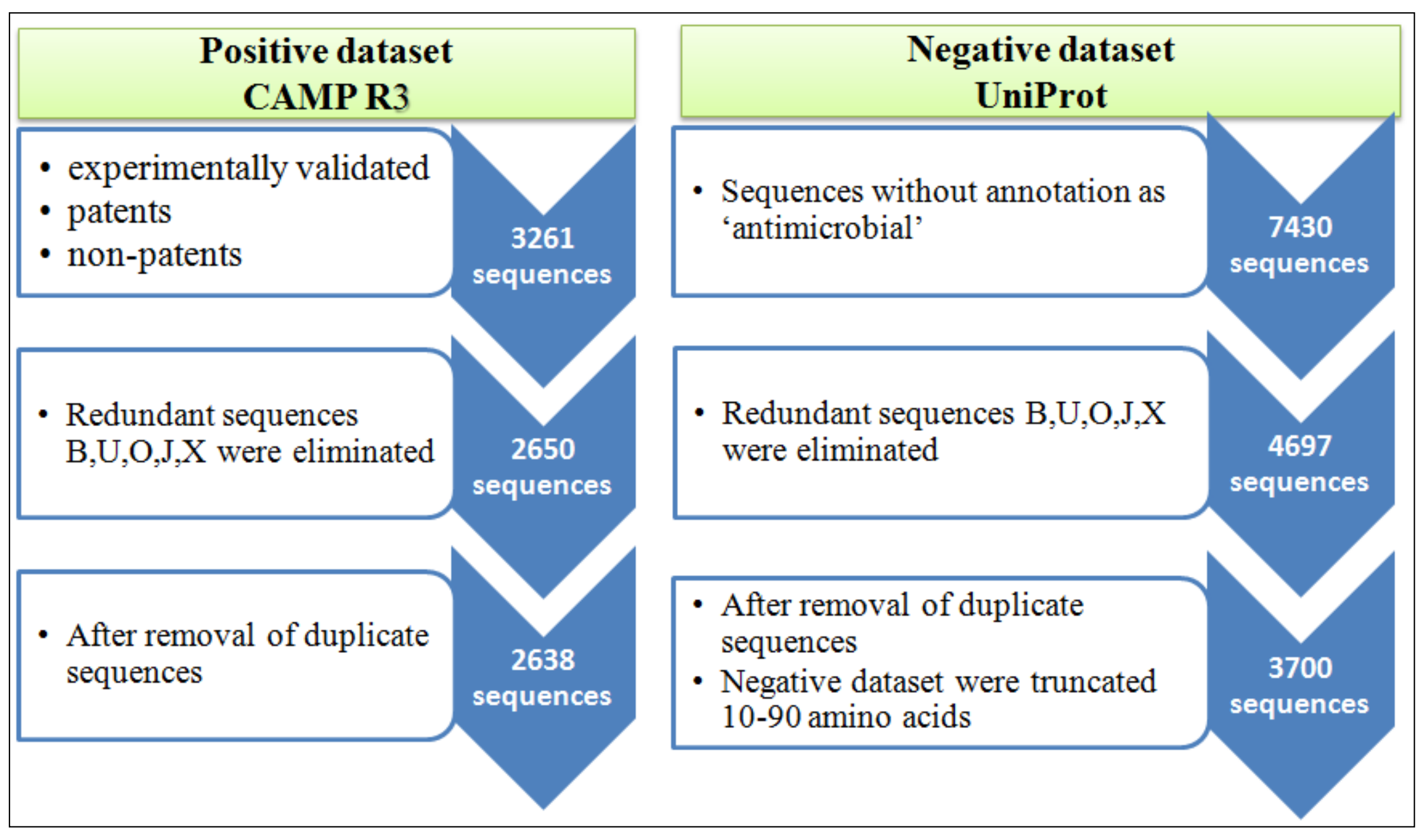
K-fold cross validation like 5-fold and 10-fold cross validation

AMP Prediction

Prediction of antimicrobial peptides using developed model

Methodology

1. Data collection



Positive dataset CAMP R3

- experimentally validated
- patents
- non-patents

3261 sequences

- Redundant sequences B,U,O,J,X were eliminated

2650 sequences

- After removal of duplicate sequences

2638 sequences

Negative dataset UniProt

- Sequences without annotation as ‘antimicrobial’

7430 sequences

- Redundant sequences B,U,O,J,X were eliminated

4697 sequences

- After removal of duplicate sequences
- Negative dataset were truncated 10-90 amino acids

3700 sequences

❑ Finally, a dataset of 6338 peptide sequence obtained from adding negative and positive dataset were used for the model development.

2. Descriptor calculation

❑ Descriptors of the retrieved sequences were calculated by using open source chemo- informatics toolkit PyDPI-1.0. It calculated 1067 protein descriptors.

Table 1. Descriptors computed for protein sequences.

Sr. No.	Description	Type	Number
1	Amino acid composition	AA Comp	20
2	Dipeptide composition	DP Comp	400
3	Geary autocorrelation Descriptors	Geary Auto	240
4	Moran autocorrelation Descriptors	Moran Auto	240
5	Composition, transition and distribution	CTD	147
6	Pseudo amino acid composition at (λ = 0)	(Psc-AAC)	20
Total Sequence based descriptors		1067	

3. Feature selection

❑ In this work firstly normalization was performed by MinMaxScaler method of Scikit-learn (0.18.1) to scale randomized data, so that they could be compared relevantly.

❑First the training dataset was scaled in the range of 0 to 1 and then test dataset was scaled by using the training dataset.

❑Feature selection was performed by using Random Forest Classifier in Scikit-learn (0.18.1).

❑Descriptors were selected for model development

❑ Out of 1067 descriptors, 824 descriptors were selected randomly on the basis of variance threshold at the value of 0.003.

❑Out of these 824 descriptors Random Forest Classifier was applied and 54 descriptors were selected for model development

4. Model development

❑ The Scikit-learn Machine Learning in Python tool kit was used for the development of model.

❑ The Random Forest, SVM and *k*-NN machine learning techniques were used to develop the model.

❑ The whole dataset was divided into training and test set into 4:1 ratio.

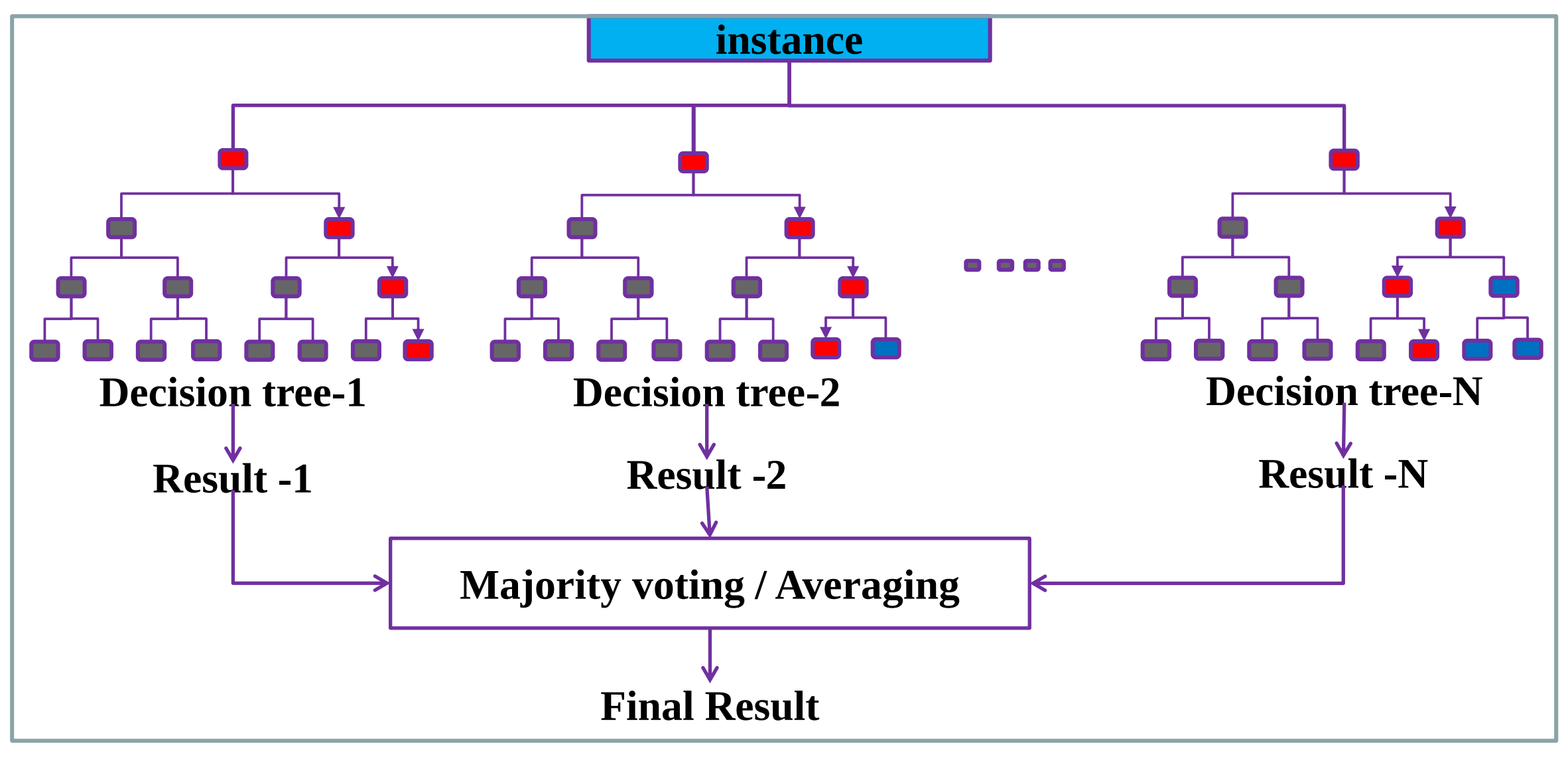
❑ Firstly, the model was trained on the training set by the use of previously selected 54 descriptors.

❑The trained model was used to predict the test set for checking the Accuracy, specificity, sensitivity, Precision, and F1-score

4.1. Random forest:

❑ Random forest is an ensemble classifier that consists of many decision trees and outputs the class that is the mode of the class output by individual trees

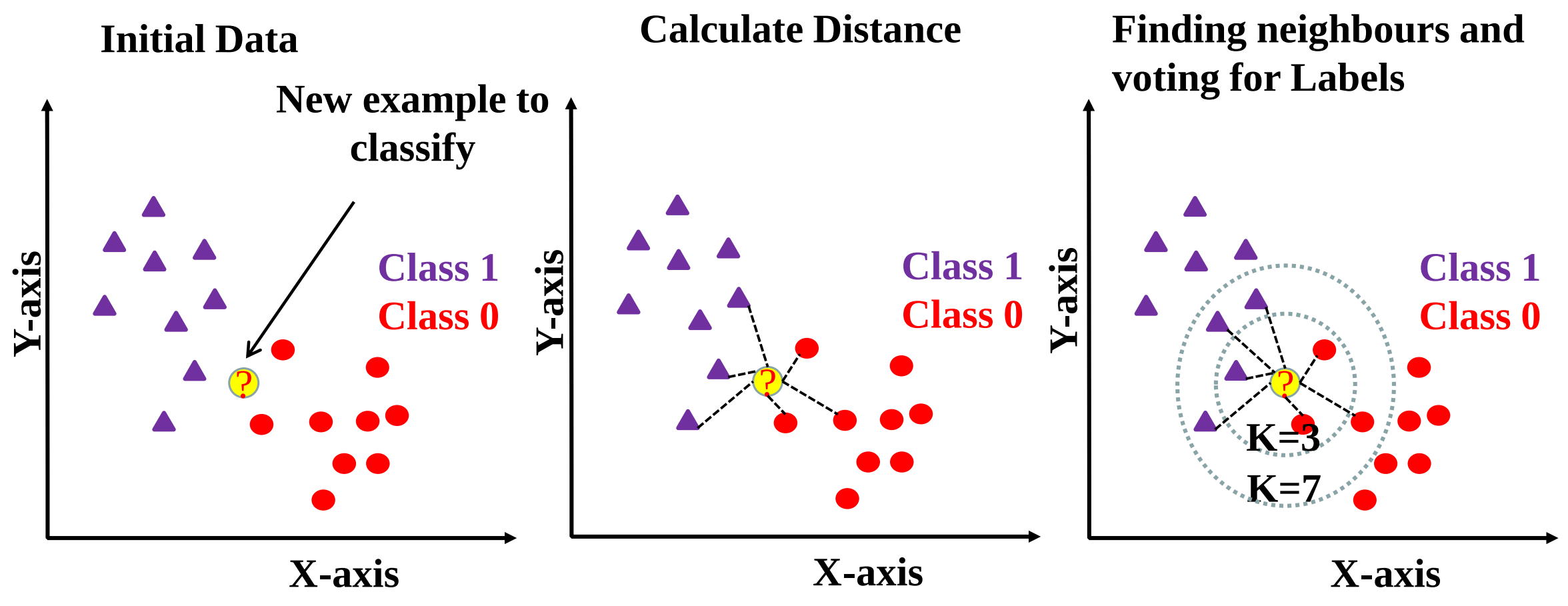
❑ Decision trees are individual learners that are combined. They are one of the most popular learning methods commonly used for data exploration



4.2. *k*-Nearest Neighbor (k-NN):

❑ Is a non-parametric method used for classification and regression. In both cases, the input consists of the *k* closest training examples in the feature space.

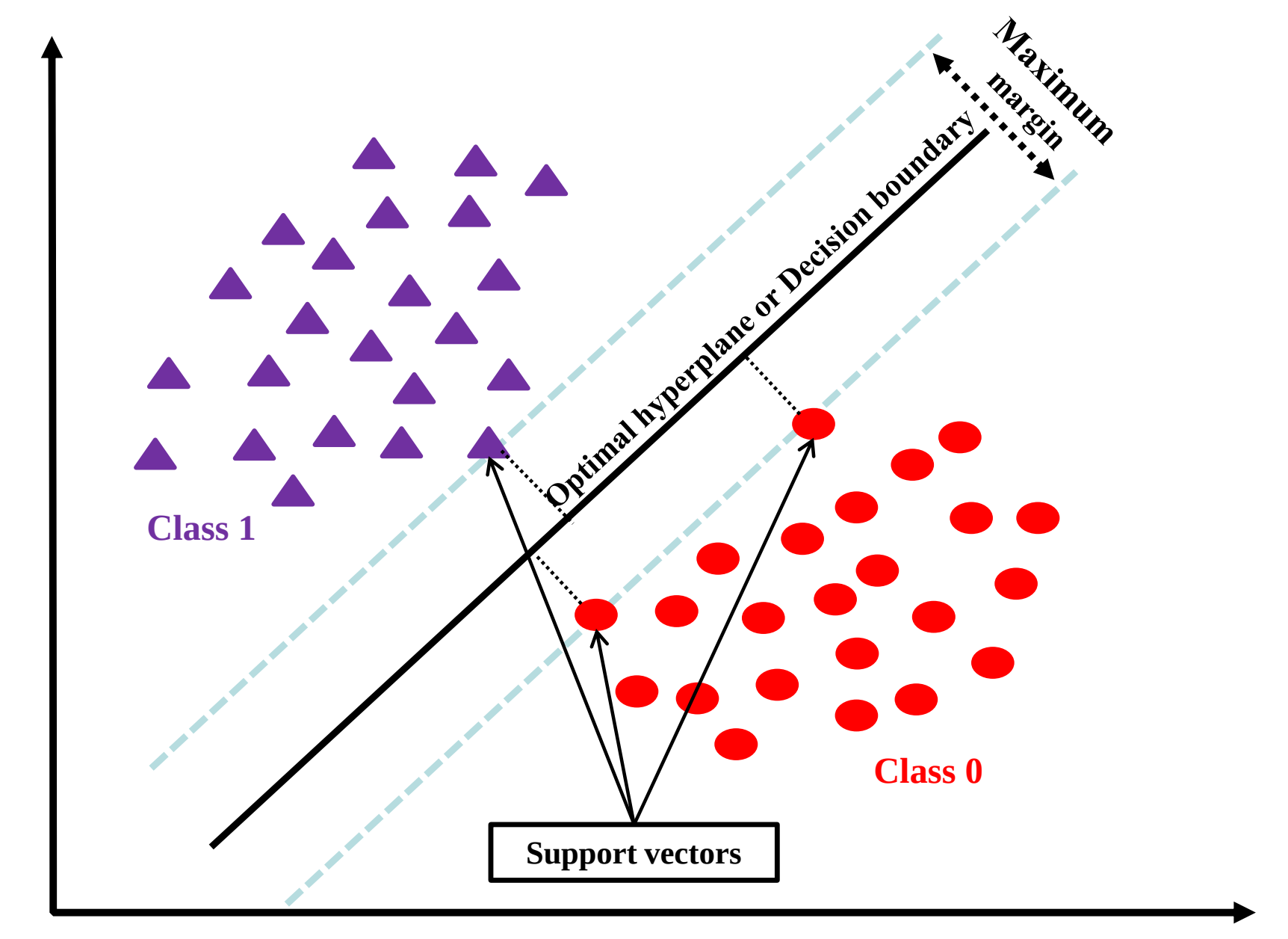
❑ The degree of smoothing is controlled by k, the number of neighbors taken into account, which is much smaller than N, the sample size



4.3. Support Vector Machine (SVM):

❑ Support vector machine is supervised learning models with associated learning algorithms that analyze data used for classification

❑ SVM training algorithm builds a model that assigns new examples into one category or the other, making it a non-probabilistic binary linear classifier



5. Validation (Enrichment Analysis)

❑ The correctness of a classification can be evaluated by computing the number of correctly recognized class examples (true positives, TP), the number of correctly recognized examples that do not belong to the class (true negatives, TN), and examples that either were incorrectly assigned to the class (false positives, FP) or that were not recognized as class examples (false negatives, FN).

Table 2. Parameters used in enrichment analysis.

Parameter	Formula	Definition
Accuracy	(TP+TN)/(TP+TN+FP+FN)	The percentage of predictions that are correct
Precision	TP/(TP+FP)	The percentage of positive predictions that are correct
Sensitivity	TP/(TP+FN)	The percentage of positive instances that were predicted as positive
Specificity	TN/(TN+FP)	The percentage of negative instances that were predicted as negative

Results

Table 3. Statistical parameters for 5-fold cross validation and external validation for RF model at threshold 0.997

n_est	Dataset	Accuracy	Precision	Specificity	Sensitivity	f1_score	1s	0s
10	Internal CV	0.8943	0.8892	0.9243	0.9143	0.8703	2110	2960
	External CV	0.8904	0.8762	0.9135	0.8935	0.867	528	740
25	Internal CV	0.9055	0.8823	0.9152	0.9052	0.8871	2110	2960
	External CV	0.9046	0.879	0.9122	0.9022	0.8864	528	740
75	Internal CV	0.914	0.8937	0.9236	0.8936	0.8971	2110	2960
	External CV	0.9046	0.879	0.9122	0.9122	0.8864	528	740

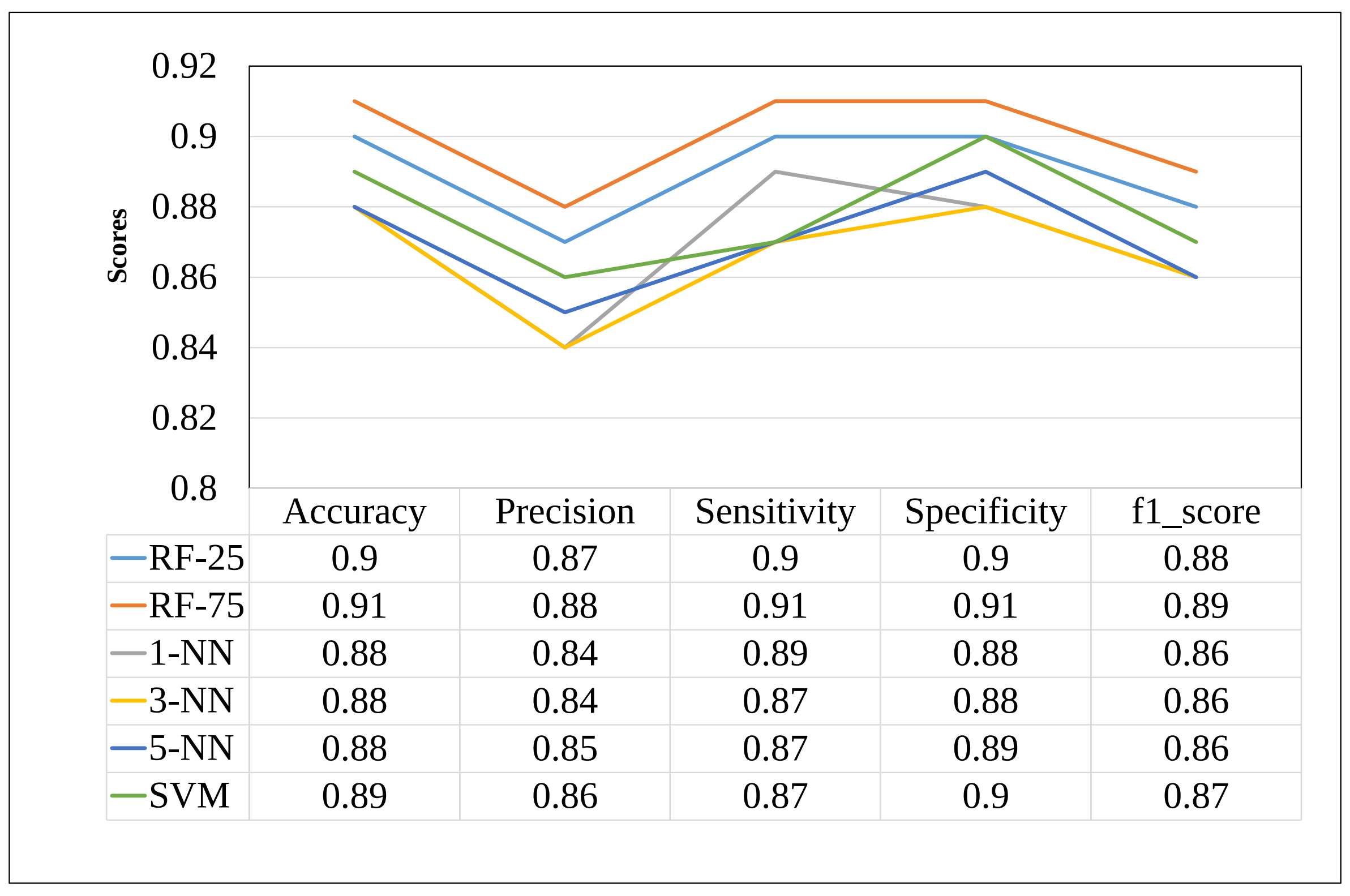
Table 4. Statistical parameters for 5-fold cross validation and external validation for *k-Nearest Neighbour* model at threshold 0.997

n_neigh	Dataset	Accuracy	Precision	Specificity	Sensitivity	f1_score	1s	0s
1	Internal CV	0.9026	0.8919	0.9247	0.9101	0.8816	2110	2960
	External CV	0.8959	0.8708	0.9068	0.9121	0.8757	528	740
3	Internal CV	0.8964	0.8925	0.9267	0.8729	0.8729	2110	2960
	External CV	0.8983	0.8859	0.9203	0.8766	0.8766	528	740
5	Internal CV	0.8964	0.8914	0.9257	0.873	0.873	2110	2960
	External CV	0.8951	0.8835	0.9189	0.8725	0.8725	528	740

Table 5. Statistical parameters for 5-fold cross validation and external validation for Support Vector Machine model at threshold 0.997

Dataset	Accuracy	Precision	Specificity	Sensitivity	F1_Score	1s	0s
Internal CV	0.8978	0.8812	0.9162	0.8100	0.8766	2110	2960
External CV	0.8904	0.8649	0.9027	0.8404	0.8690	528	740

Performance of classifiers



	Accuracy	Precision	Sensitivity	Specificity	f1_score
RF-25	0.9	0.87	0.9	0.9	0.88
RF-75	0.91	0.88	0.91	0.91	0.89
1-NN	0.88	0.84	0.89	0.88	0.86
3-NN	0.88	0.84	0.87	0.88	0.86
5-NN	0.88	0.85	0.87	0.89	0.86
SVM	0.89	0.86	0.87	0.9	0.87

External Validation: Sequences, with known functions, were predicted by using the developed model through the command line interface. The model predicted with 100% accuracy (Table 6).

Table 6. External validation results.

Peptide_ID	Title	Sequence	Activity
CAMPSQ7	Sesquin	KTCENLADTY	AMP
CAMPSQ17	Antifungal lectin PVAP	SNDIYFNFQR	AMP
CAMPSQ519	Gymnin	KTCENLADDY	AMP
CAMPSQ593	Antimicrobial ribonuclease	DNGEAGRAAR	AMP
P29135	Neurokinin-A	HKLDSFIGLM	Non-AMP
P86303	Tachykinin-like peptide-VI (PnlTkP-VI)	QKKDRFLGLM	Non-AMP
P86305	Tachykinin-like peptide-VIII (PnlTkP-VIII)	QKKDKKDRFY	Non-AMP
P16224	Locustatachykinin-2 (Locustatachykinin II) (TK-II)	APLSGFYGVR	Non-AMP
P08608	Scyliorhinin-1 (Scyliorhinin I)	AKFDKFYGLM	Non-AMP

Conclusion

❑ Several machine learning models were developed using different classifiers like Random forest, k-NN and SVM for the prediction of antimicrobials and non-antimicrobials, among all RF has shown good results comparatively

❑ From assessment of the classifiers, it is concluded that random forest with (n_estimators) = 75 shows good performance as compared to other classifiers

❑Unknown sequences were successfully predicted by developed model

❑**Wani, M.A.,**P Garg,. Roy, K.K. Machine learning-enabled predictive modeling to precisely identify the antimicrobial peptides. Medical & Biological Engineering & Computing (2021).

Acknowledgements

❑ The authors express their gratitude to the Director of NIPER Mohali and Ministry of Chemicals and fertilizes for providing fellowship.

References

1. Sang, Y.; Blecha, F., Antimicrobial peptides and bacteriocins: alternatives to traditional antibiotics. *Anim. Health. Res. Rev.* ***2008, 9,*** ***227-235.***
2. Waghu, F. H.; Barai, R. S.; Gurung, P.; Idicula-Thomas, S., CAMPR3: a database on sequences, structures and signatures of antimicrobial peptides. *Nucleic acids res.* ***2015, 1051.***
3. Hammami, R.; Hamida, J. B.; Vergoten, G.; Fliss, I., PhytAMP: a database dedicated to antimicrobial plant peptides. *Nucleic acids res.* ***2009, 37, D963-D968.***
4. Waghu, F. H.; Gopi, L.; Barai, R. S.; Ramteke, P.; Nizami, B.; Idicula-Thomas, S., CAMP: Collection of sequences and structures of antimicrobial peptides. *Nucleic acids res.* ***2014, 42, D1154-D1158.***