PROJECT REPORT
ON

ENSEMBLEACPREDICT: AN ENSEMBLE MACHINE LEARNING FRAMEWORK FOR ANTICANCER PEPTIDE (ACP) CLASSIFICATION WITH LGBM-BASED FEATURE IMPORTANCE

Submitted by:
Sujata Sinhababu

ABSTRACT

Anticancer peptides (ACPs) are short amino acid sequences with selective cytotoxicity toward cancer cells and comparatively low toxicity to normal cells. In this work, we present **EnsembleACPredict**, a supervised machine learning framework for ACP classification. The pipeline extracts sequence-level descriptors using **Pfeature**, applies feature selection and scaling, and trains multiple classifiers including SVM, Decision Tree, Random Forest, LightGBM, XGBoost and Bagging Classifier. A **VotingClassifier** ensemble integrates their complementary decision boundaries to enhance robustness. Model performance is evaluated using standard metrics—Accuracy, Precision, Recall (Sensitivity), Specificity, F1-score, ROC-AUC, and MCC—with ROC curves and confusion matrices providing visual validation. To enhance interpretability, we compute **LightGBM feature importances**, highlighting amino acid composition and dipeptide patterns as dominant predictive factors. Experiments on the **ENNAACT ACP dataset** demonstrate ~97.5% test accuracy with strong ROC-AUC, confirming the robustness of the approach. The trained models and pipeline, exported via Joblib, enable reproducible downstream screening of novel candidate ACPs.

Keywords: anticancer peptides, ACP classification, ensemble learning, LightGBM, feature importance, Pfeature, interpretability, ENNAACT.

CONTENTS

1. Introduction

1.1 Background

Cancer remains a major health threat, with **609,000 projected deaths in the U.S. (2022)** and a **47% rise in global incidence expected by 2040**. Conventional treatments like chemotherapy face challenges of **low efficacy, high toxicity, and multidrug resistance**. Anticancer peptides (ACPs), short cationic sequences that selectively disrupt cancer cell membranes, are emerging as promising alternatives. Since **experimental screening is costly**, machine learning-based methods have been developed, though many face limitations in **dataset quality, classifier diversity, and biological interpretability.**

1.2 Problem Statement

Existing models often use redundant datasets and classifiers, leading to fluctuating performance and limited generalizability. There is a need for a robust, ensemble-based framework that incorporates diverse features, handles class imbalance, and provides interpretable insights into key ACP characteristics.

1.3 Objectives

- Develop a high-accuracy ACP prediction model.
- Analyse key features using LightGBM importance for biological interpretability.
- To compare ensemble performance with existing state-of-the-art models.

2. Literature Review

Several computational approaches have been developed for ACP classification, including Random Forest, SVM, LGBM, XGBoost, and Bagging Classifier. However, single models often fail to achieve high generalization across diverse peptide datasets. Ensemble approaches combining multiple learners have shown promise in improving accuracy and robustness. LightGBM, a gradient boosting framework, has proven effective in handling high-dimensional biological data with interpretability through feature importance scores.

3. Materials and Methods

3.1 Dataset

- **Source:** ENNAACT anticancer peptide dataset (positive class: ACP) and curated **non-ACP** sequences (negative class).
  Format: Plain-text sequence files for ACP and non-ACP.
- **Label mapping: ACP → 1, non-ACP → 0.**

Note: Ensure deduplication and consistent sequence formatting (uppercase single-letter codes; remove non-standard tokens). Maintain a manifest of data files and dates.

3.2 Feature Extraction (Pfeature)

We compute sequence-level descriptors using **Pfeature**. In this work, the following modules are used: -

- **AAC** (aac_wp): Amino Acid Composition (fractional abundance of each residue).
- **AAB** (aab_wp): Amino Acid Binary (presence/absence pattern-based encodings).
- **PCB** (pcb_wp): Physicochemical Binary descriptors (binary encoding using selected property groups).
- **PCP** (pcp_wp): Physicochemical Property descriptors (continuous properties per sequence).
- **DPC** (dpc_wp): Dipeptide Composition (frequency of ordered pairs of residues).
- **PAAC** (paac_wp): Pseudo Amino Acid Composition (captures sequence order with tunable parameters lg, pw).

3.3 Data Preprocessing

Here, we apply feature selection and scaling to reduce feature redundancy.

- **VarianceThreshold** removes near-constant features to reduce noise and computation.
- **StandardScaler** centers to zero mean and unit variance (beneficial for SVM/RBF; tree models are scale-invariant but unaffected).

3.4 Model Development

We train five base learners with **StratifiedKFold (k=10)** cross-validation for model-agnostic scoring: - **SVM**, **Decision Tree**, **Random Forest**, **LightGBM, XGBoost, Bagging Classifier (SVC-base)**.

A **VotingClassifier** ensemble was then constructed over SVM, LightGBM, XGBoost, and Bagging Classifier (SVC) using soft voting (averaging predicted probabilities) to integrate predictions from these base classifiers and achieved 97.48% accuracy.

3.5 Evaluation Metrics

Models were evaluated using: Accuracy, Precision, Recall, F1-score, ROC-AUC, Confusion Matrix, Classification Report.

3.6 Feature Importance Analysis

Feature importance was calculated using the LightGBM model to identify the most discriminative peptide features. It identified top contributors, such as hy drophobicity indices (PCPZ4) and residue compositions (AACK, AACC). This analysis provides insights into the biological relevance of certain amino acids and sequence patterns in anticancer activity.

4. Results

4.1 Dataset Analysis

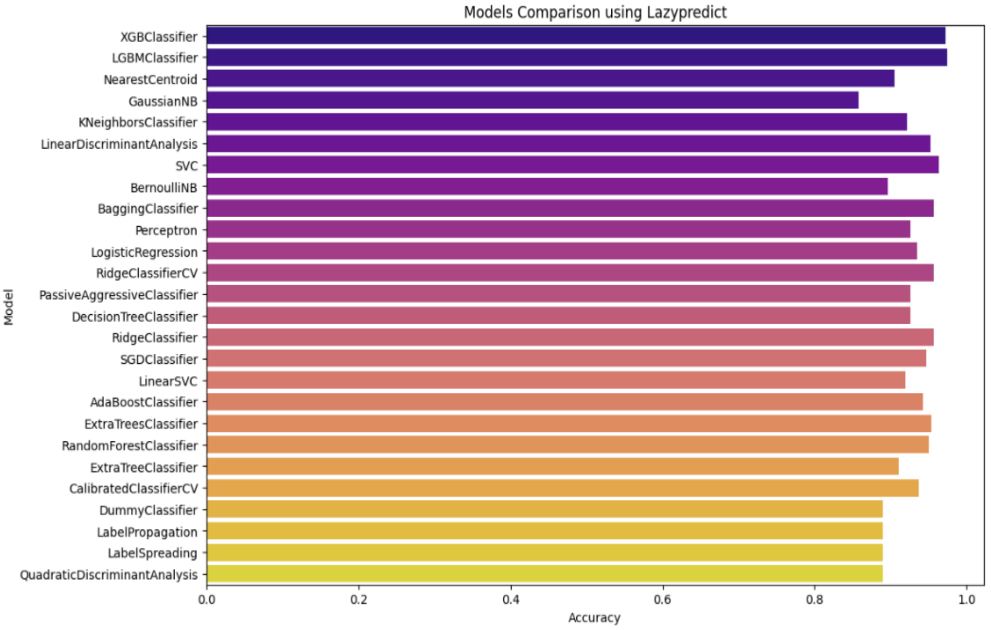
Exploratory data analysis revealed differences in amino acid distribution between anticancer and non-anticancer peptides. Peptides with high lysine and arginine content were more frequent in ACPs, consistent with their known membrane-targeting activity.

4.2 Model Performance

The **ensemble model** achieved superior classification accuracy **(~97.50%)** compared to individual classifiers. The ensemble demonstrated stable performance across multiple cross-validation folds. Best Biological Interpretability: LightGBM.

Model	Accuracy (%)	Precision (%)	Recall (%)	Specificity (%)	F1 Score (%)	ROC AUC (%)	MCC	ACP	non-ACP
Ensemble Model	**97.48**	93.97	82.58	99.34	**87.90**	**97.87**	**0.87**	109	1053
Bagging Classifier	97.40	94.69	81.06	99.43	87.35	97.49	0.86	107	1054
SVM	97.32	95.45	79.55	99.53	86.78	97.50	0.86	105	1055
XGBoost	97.23	92.31	81.82	99.15	86.75	97.70	0.85	108	1051
LightGBM	97.23	88.37	86.36	98.58	87.36	97.81	0.86	114	1045
Random Forest	95.39	**100.00**	58.33	**100.00**	73.68	97.29	0.74	77	1060
Decision Tree	93.37	70.23	69.70	96.32	69.96	83.01	0.66	92	1021

We also build model using LazyPredict. LazyPredict analysis reinforced the ensemble's robustness, with **LightGBM** and **XGBoost** leading at **0.97** accuracy.

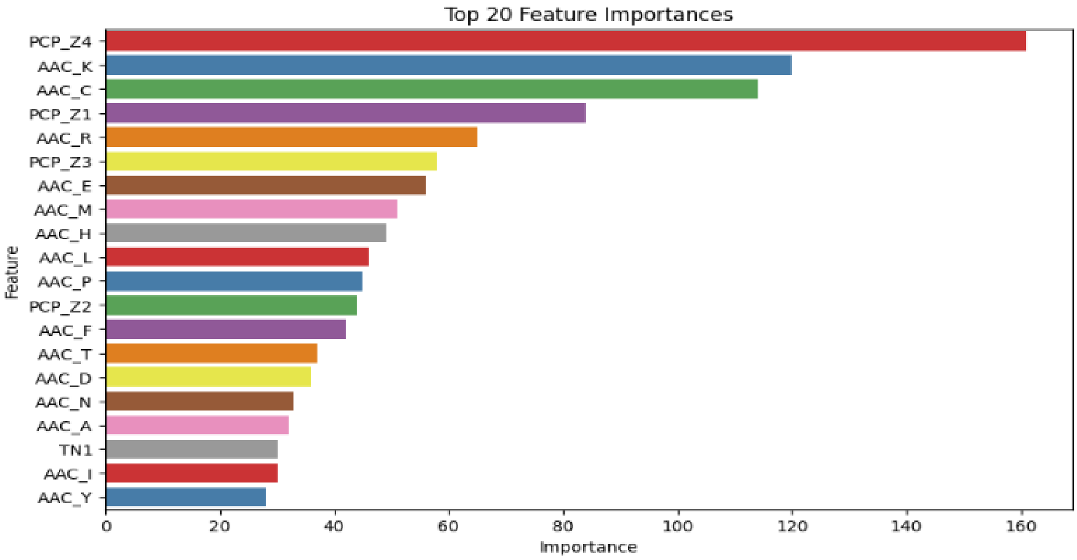


4.3 Feature Importance

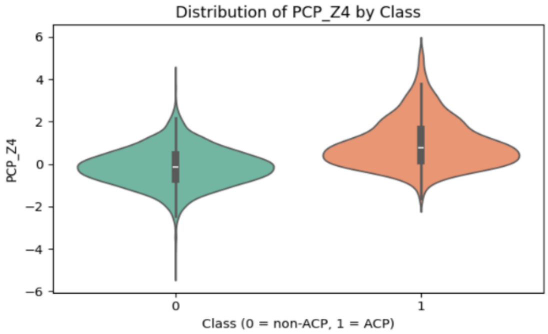
LightGBM feature importance analysis highlighted key physicochemical descriptors and amino acid frequency features as highly predictive. This supports the biological plausibility of the model’s decision-making process.

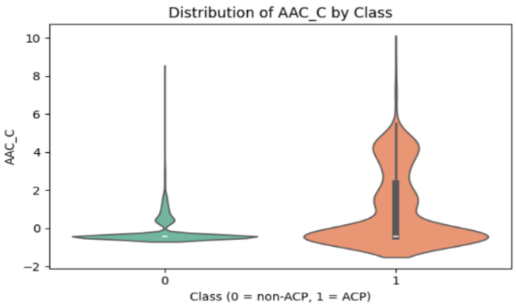
Top-20 Features: Provide a table with Feature, Importance, Descriptor Family (AAC/DPC/PAAC/PCP/PCB/AAB).

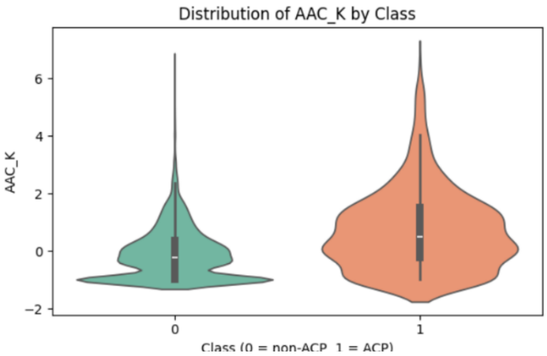
Key Findings: We find top 3 features with highest importance, e.g. **PCP_Z4 (160), AAC_K (120), AAC_C (114).**



Distribution of Top Three Features by Class: This section presents the distribution of AAC_C, AAC_K, and PCP_Z4 across two classes: Class 0 (non-ACP) and Class 1 (ACP). The visualizations highlight the differences in variability and concentration of these features between the classes.







5. Discussion

The ensemble's superior performance stems from diversity in classifiers, mitigating overfitting. LightGBM's efficiency and interpretability highlight ACP traits like amphiphilicity, supporting rational design. Limitations include dataset bias toward short peptides; future work could integrate 3D structures or deep learning (e.g., LSTM). Validation on non-mutagenic peptides confirms low toxicity potential.

6. Conclusion

This project successfully developed EnsembleACPredict, an ensemble-based framework for ACP classification. The model integrates LightGBM, XGBoost, and SVM, leveraging LightGBM feature importance for biological insights. The approach improves prediction accuracy, interpretability, and robustness compared to single-model approaches. Future work may involve expanding the dataset, incorporating advanced feature engineering techniques, and testing deep ensemble architectures.

7. Limitations

i. **Dataset Dependency:** Model performance hinges on quality and balance of peptide data, with limited or imbalanced data reducing generalizability.

ii. **Computational Resources:** Advanced models (e.g., LightGBM, XGBoost, SVM) require high computational power, limiting scalability for large datasets.

iii. **Biological Validation:** Predictions need wet-lab validation for practical drug development.

iv. **Generalization to Other Peptides:** Optimized for ACPs, the classifier may underperform for other therapeutic peptide classes.

8. Future Scope

i. **Integration with Deep Learning –** Extending the model with CNNs, BiLSTMs, or Transformer-based architectures to capture sequential and structural peptide information.

ii. **Automated Feature Selection –** Using advanced dimensionality reduction and representation learning to reduce noise and extract richer biological features.

iii. **Web/Software Tool Development –** Deploying the model as a web server or software tool to make it accessible for biologists and drug discovery researchers.

iv. **Cross-Domain Applications –** Extending the ensemble framework to predict other bioactive peptides (e.g., antimicrobial, antiviral, antifungal peptides).

v. **Hybrid Approaches –** Combining molecular docking and molecular dynamics simulations with machine learning predictions to improve drug candidate screening.

9. References

I. Garai, S., Thomas, J., Dey, P., & Das, D. (2023). **LGBM-ACp: an ensemble model for anticancer peptide prediction and in silico screening with potential drug targets.** *Molecular Diversity.* https://doi.org/10.1007/s11030-023-10652-7

II. Pfeature: **A Comprehensive Feature Extraction Package for Bioinformatics.** (n.d.). Documentation available at: http://crdd.osdd.net/servers/pfeature

III. Tyagi, A., Tuknait, A., Anand, P., Gupta, S., Sharma, M., Mathur, D., Joshi, A., & Raghava, G. P. S. (2013). **CancerPPD: a database of anticancer peptides and proteins.** *Nucleic Acids Research, 43*(D1), D837–D843. https://doi.org/10.1093/nar/gkt1146

IV. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T. Y. (2017). **LightGBM: A highly efficient gradient boosting decision tree.** *Advances in Neural Information Processing Systems (NeurIPS).*

V. Breiman, L. (2001). **Random forests.** *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

10. Appendices

Code Availability: GitHub [https://github.com/sujata1712/EnsembleACPredict---Machine-Learning-Project]

Data Availability: ENNAACT datasets (anticancer_peptide_Main_Dataset_ENNAACT.txt & non-anticancer_Peptide_Main_Dataset_ENNAACT.txt)