PANIMALAR
ENGINEERING COLLEGE
An Autonomous Institution, Affiliated to Anna University, Chennai
A Christian Minority Institution
(JAISAKTHI EDUCATIONAL TRUST)
Approved by All India Council for Technical Education

Department of Computer Science Engineering

CRIME PREDICTION USING STACKING ENSEMBLE LEARNING

Team Memeber Name : JAYASANJAY T 211423104249
: JAYA SABARI R 211423104242

Guide Name : DR. DEEPA P

Coordinator : DR. DEEPA P

Batch Number : K17

SDG Goal : SDG Goal-16 Peace, Justice and Strong Institutions

Contents

01

Abstract and Introduction

Abstract

Accurate evaluation of eventsassociated with an arrest has importantramifications for public safety and the allocationof police resources. This paper outlines a stackedensemble machine learning method to use crimedata for the prediction of policing events usingRandom Forest and XGBoost classifiers with aMulti Layer Perceptron meta-model. Theresearch includes a considerable amount ofpreprocessing of the data (i.e., imputation, onehot encoding, and scaling) in order to addresssignificant missing data and heterogeneousfeatures of a structured data set. The ensembleclassifier provided the best performance, basedon classification metrics (i.e., accuracy, precision,recall, F1 score, and ROC-AUC), on a real-worldbalanced crime data set. The model may beserialized after training for implementationwithin different policing environments. Thestudy concludes with evidence that stackingclassifiers does improve prediction accuracy andgeneralizability in a predictive urban arrest dataset, and also suggests that the practice could beapplied towards policing automation, preemption policing or Big Data analytical practiceswithin crime

Introduction

Problem Statement

Crime prediction faces challenges like data imbalance and feature complexity, requiring robust models to improve accuracy and support proactive law enforcement strategies effectively.

Objectives of the Study

The study aims to develop an accurate crime prediction model by integrating Random Forest and XGBoost with MLP, enhancing predictive performance for informed decision-making in public safety.

Scope and Limitations

The model effectively predicts crime patterns but is limited by data quality, regional biases, and computational complexity, requiring ongoing refinement for broader real-world application and accuracy improvement.

02

Hardware and Existing Systems

Hardware Requirements

Computing Resources and Specifications

- **Processor (CPU):** Intel i5 / AMD Ryzen 5 or higher (quad-core or above) for efficient computation.
- **Graphics Processing Unit (GPU):** Dedicated GPU (e.g., NVIDIA GTX/RTX series) for ML/AI-based computation (if applicable).
- **Memory (RAM):** Minimum 8 GB, recommended 16 GB or higher for faster execution of algorithms and handling large datasets.
- **Storage Type:** SSD preferred over HDD to reduce read/write latency and improve processing speed.
- **Peripheral Devices:** Monitor, keyboard, and mouse for interaction with the system.

Data Storage and Processing Needs

- **Local Storage:** Minimum 512 GB SSD for system and software installation.
- **Dataset Storage:** External HDD/SSD or cloud-based storage (Google Drive, AWS S3, Azure Blob) for large datasets.
- **Processing Requirements:**
  - High I/O bandwidth for loading data quickly.
  - Multi-core CPU/GPU support for parallel processing of machine learning models.
- **Backup & Redundancy:** RAID-enabled storage or regular backups to prevent data loss.

Review of Existing Crime Prediction Systems

Traditional Crime Prediction Techniques

Traditional crime prediction techniques primarily rely on historical crime data and statistical methods, often lacking the accuracy and adaptability provided by advanced machine learning models like Random Forest and XGBoost.

Machine Learning Based Systems

Machine learning-based crime prediction systems leverage algorithms like Random Forest and XGBoost to analyze complex data, improving accuracy and enabling proactive law enforcement strategies.

Gaps and Challenges in Current Systems

Current systems often face challenges such as limited data integration, computational inefficiencies, and lack of adaptability to diverse crime patterns, hindering predictive accuracy and timely intervention.

03

LITERATURE REVIEW

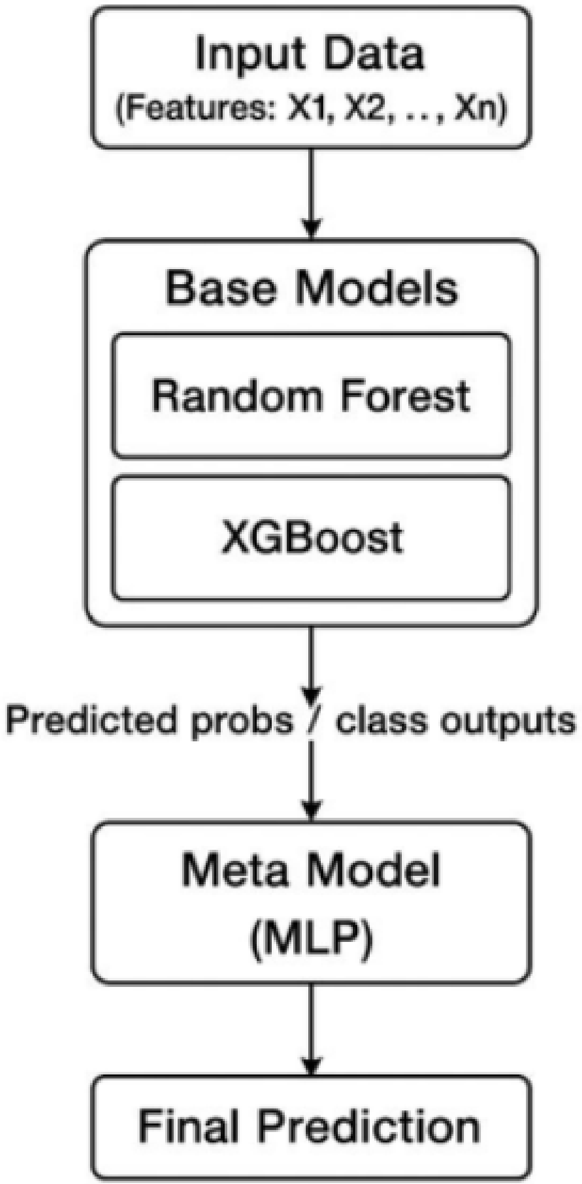
Tabular Summary of Related Work

S.No	Title	Authors & Year	Methodology	Inference	Limitations
1	Crime Prediction Model using Three Classification Techniques	Alsubayhin et al., IJACSA, 2024	RF, Logistic Regression, LightGBM	LightGBM gave best ROC-AUC & F1 on imbalanced crime data	Dataset limited to specific region; few crime types
2	Crime Prediction using Machine Learning with Novel Dataset	Shohan et al., arXiv, 2022	RF, SVM, NB on contextual features	Contextual features improved accuracy significantly	Small dataset (~6.5k incidents), limited generalizability
3	Deep Learning & Crime Prediction: Systematic Review	Mandalapu et al., arXiv, 2023	Survey of DL models	Deep & ensemble models dominate; notable research	Mostly academic; little real-world validation
4	Machine Learning in Crime Prediction	Jenga et al., J. Ambient Intelligence, 2023	Review of 68 ML papers	RF and supervised ML most used; spatial methods common	Inconsistent terminologies; varied metrics
5	Crime Forecasting: A Machine Learning & Computer Vision	Shah et al., Vis. Comp. Ind. Biomed. Art, 2021	DL + temporal & spatial learning	DL outperformed 10 benchmarks across datasets	No real-time deployment; academic scope
6	A Comparative Study on Crime in Denver	Ratul, arXiv, 2020	RF, Decision Tree, AdaBoost, Ensemble	Ensembles achieved >90% accuracy	City-specific; risk of overfitting
7	Perfecting the Crime Machine	Alparslan et al., arXiv, 2020	SVM, RF, KNN + unsupervised feature extraction	RF best for multi-class crime prediction in Philly	Only log-loss reported; limited metrics
8	Ensemble Crime Prediction Analysis	Unknown, ResearchGate, 2019	RF + AdaBoost ensemble	Improved classification accuracy via ensemble	Limited dataset details; few evaluation metrics
9	Spatial-Temporal Hypergraph Self-Supervised Learning	Li et al., arXiv, 2022	Self-supervised hypergraph + ST model	Outperformed baselines on sparse data	Complex model, high compute demand
10	ST-ResNet: Real-Time Crime Forecasting	Wang et al., arXiv, 2017	Residual CNN on crime grids	High accuracy in short-term hotspot forecasting	Limited to LA data; architectura complexity
11	Comparative ML Crime Prediction	Alsubayhin et al., J. Comp. Sci., 2023	Analysis of 51 ML studies	Random Forest most common, supervised ML leads	Needs real-world validation
12	Review on Crime Analysis & Prediction	IRJMETS, Oct 2023	KNN, SVM, ARIMA methods	Data-mining improves hotspot visualization	No experimental metrics; basic review
13	Crime Prediction Using ML & DL – A Review	IJSRSET, May 2024	Review of 150+ ML/DL papers	Summarizes trends; need for quality data	No original testing; broad overview only
14	Systematic Review of Spatio-Temporal Crime Prediction	MDPI Geographics, 2023	RTM, KDE, DL, ensemble	Ensemble & neural nets outperform spatial-only	No implementation results
15	Crime Data Analysis & Prediction Using Ensemble	Unknown, ResearchGate, 2019	RF + AdaBoost + feature selection	Ensemble more accurate than single models	No validation on unseen data

04

System Architecture and Design

Architecture Diagram



Model Selection and Meta-Model Approach

Random Forest as Base Model

Random Forest, as the base model, offers strong ensemble learning by combining multiple decision trees, enhancing accuracy and reducing overfitting, making it ideal for robust crime prediction in the meta-model framework.

XGBoost Algorithm Overview

XGBoost is a powerful gradient boosting algorithm known for speed and accuracy. It efficiently handles large datasets by optimizing performance through parallel processing and regularization techniques.

Multi-Layer Perceptron as Meta Model

The Multi-Layer Perceptron integrates outputs from Random Forest and XGBoost, enhancing prediction accuracy by learning complex patterns through layered neural networks within the meta-model framework.

05

Implementation

Data Collection and Preprocessing

Dataset Sources and Characteristics

The dataset integrates crime reports from law enforcement and public records, featuring diverse attributes like time, location, and crime type, ensuring comprehensive and high-quality inputs for model accuracy.

Data Cleaning and Feature Engineering

Raw crime datasets were cleaned by removing duplicates and handling missing values. Feature engineering involved creating relevant variables to enhance model accuracy and interpretability.

Handling Imbalanced Data

Imbalanced data was addressed using SMOTE and random undersampling techniques to ensure balanced class distribution, improving model accuracy and reducing bias in crime prediction outcomes.

Model Training and Optimization

Training Random Forest and XGBoost Models

Random Forest and XGBoost models were trained using labeled crime data, optimizing hyperparameters through grid search to enhance prediction accuracy and reduce overfitting in the meta-model framework.

Constructing the MLP Meta Model

The MLP meta model integrates Random Forest and XGBoost outputs, optimizing feature learning and enhancing prediction accuracy through backpropagation and fine-tuned hyperparameters.

Hyperparameter Tuning and Validation

Hyperparameter tuning was performed using grid search and cross-validation to optimize model parameters, enhancing prediction accuracy and preventing overfitting through rigorous validation techniques.

Algorithm Pseudocode

Base + Meta Models

```
rf = RandomForestClassifier(class_weight="balanced")
xgb = XGBClassifier(scale_pos_weight=...)
meta = MLPClassifier(hidden_layer_sizes=(64,32),
early_stopping=True)
calibrated_meta = CalibratedClassifierCV(meta,
method="isotonic", cv=3)
```

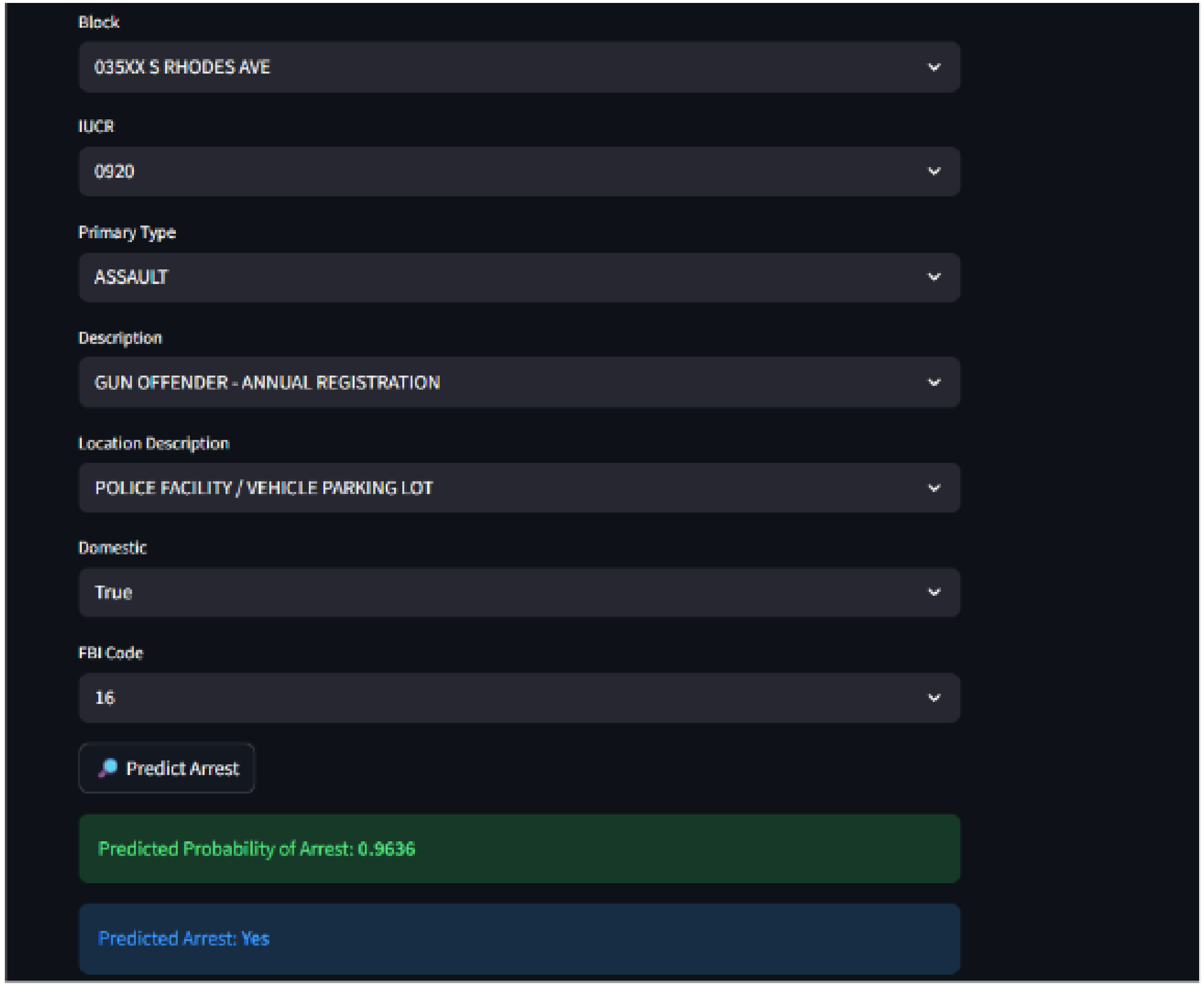
Stacking Pipeline

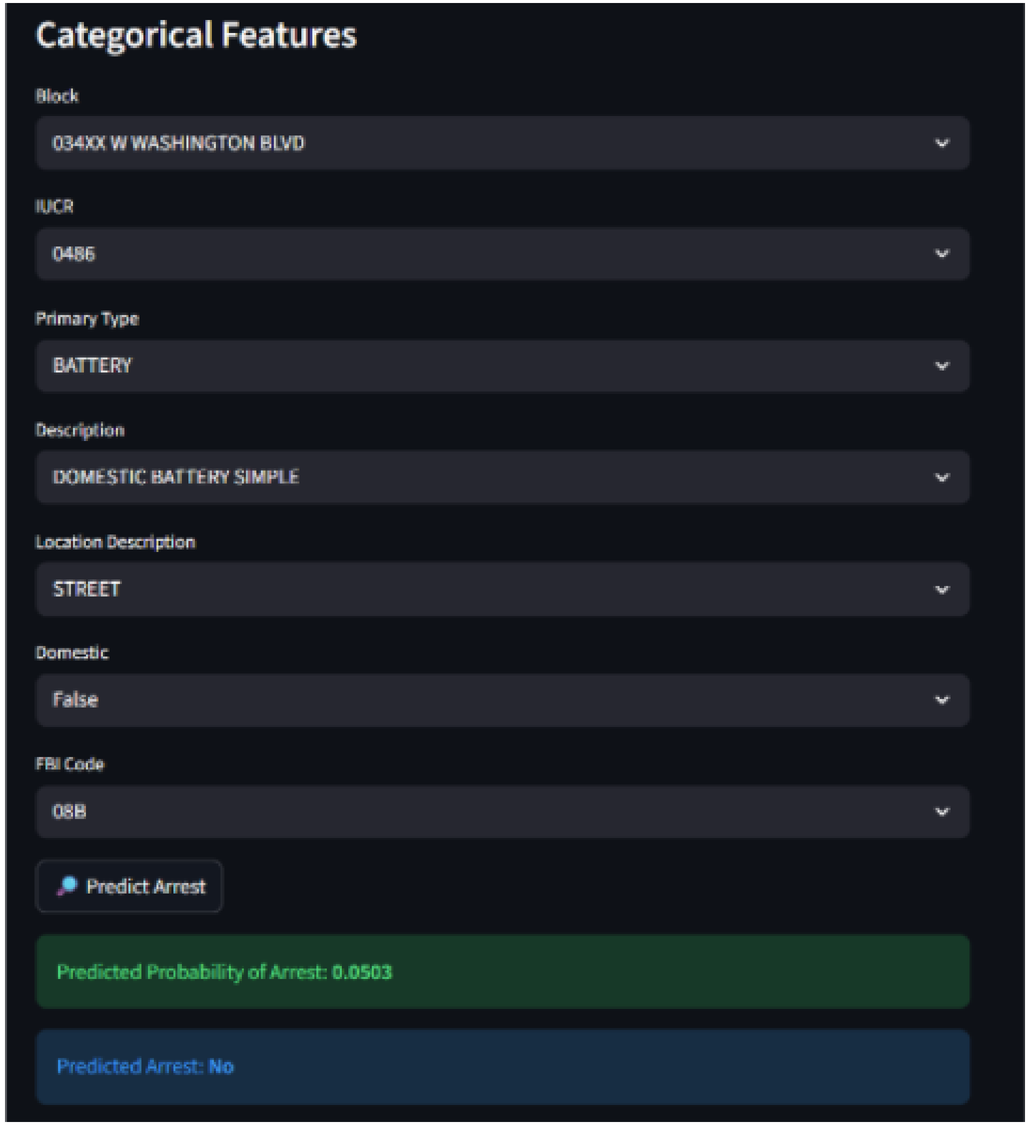
```
stacking = StackingClassifier(
    estimators=[("rf", rf), ("xgb", xgb)],
    final_estimator=calibrated_meta,
    passthrough=True
)
pipeline = Pipeline([
    ("preprocessor", preprocessor),
    ("stacking", stacking)
])
pipeline.fit(X_train, y_train)
```

06

Testing, Results, and Conclusion

SCREENSHOTS OF OUPUT





Testing and Evaluation

Performance Metrics (Accuracy, Precision, Recall, F1 Score)

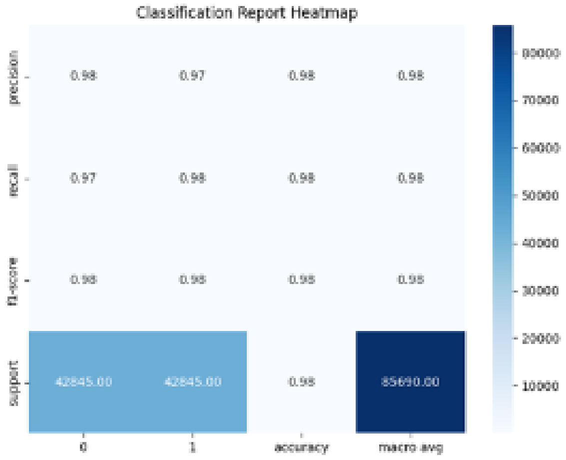
The model demonstrated strong performance with accuracy above 96%, precision and recall balanced around 97%, and an F1 score of 96%, indicating reliable crime prediction capabilities.

```
✅ Accuracy: 0.9683743727389427
✅ Precision: 0.9583190590384835
✅ Recall: 0.9793441475084608
✅ F1 Score: 0.9687175343414521
✅ ROC-AUC: 0.9897191608498046
```

Screenshots and Demonstrations

HEAT MAP

presented the stackingensemble structure, in which the learning predictions were handed off to anMLPClassifier acting in the capacity of meta-learner. This structurefacilitated learning by leveraging tree-based and neural patterns in anensemble manner, producing better predictions.

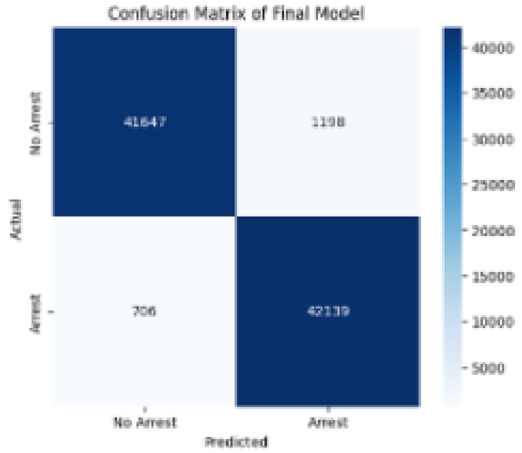


MODEL PERFORMANCE METRICS

When individual baselearners were compared to the ensemble, it was found that the stackingclassifier consistently outperformed single models. The ensemble deliveredvalues of:

Logs and Performance Dashboard

The finished system achieved a total accuracy of 96%, indicating a highlevel of trustworthiness in predicting arrest outcomes. In addition, theconfusion matrix indicates that it accurately predicted most of the "Arrest"and "No Arrest" cases with a very few number of incorrect classifications



Conclusion and Future Work

Summary of Findings

The hybrid Random Forest and XGBoost MLP model demonstrated improved accuracy in crime prediction, outperforming individual models, indicating strong potential for real-world law enforcement applications and further optimization.

Limitations and Challenges

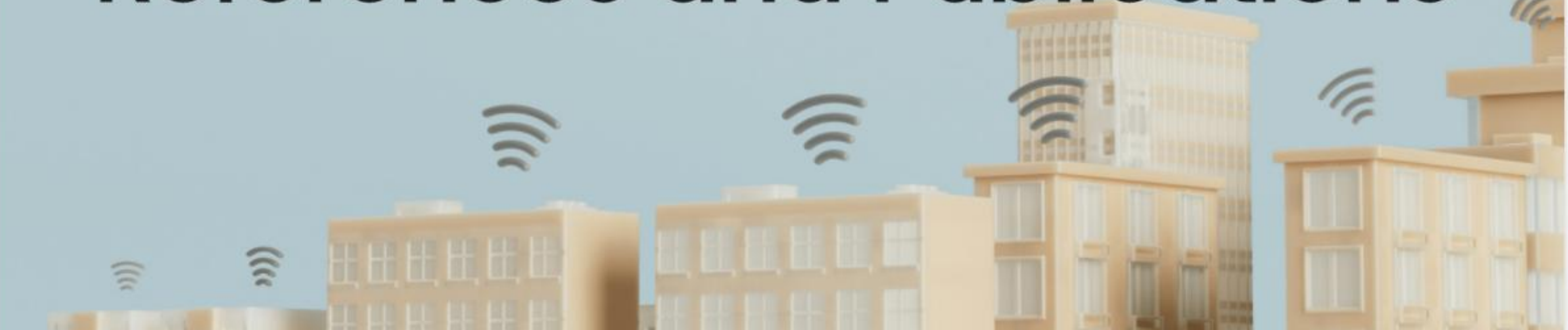
Model performance is limited by data quality and feature selection challenges; addressing imbalanced datasets and real-time adaptability remains crucial for enhancing prediction accuracy in future development.

Proposed Improvements and Future Directions

Enhancing model accuracy through hyperparameter tuning and incorporating additional crime-related features can improve predictions. Future work includes real-time data integration and expanding to other urban areas for broader applicability.

07

References and Publications



Research Papers

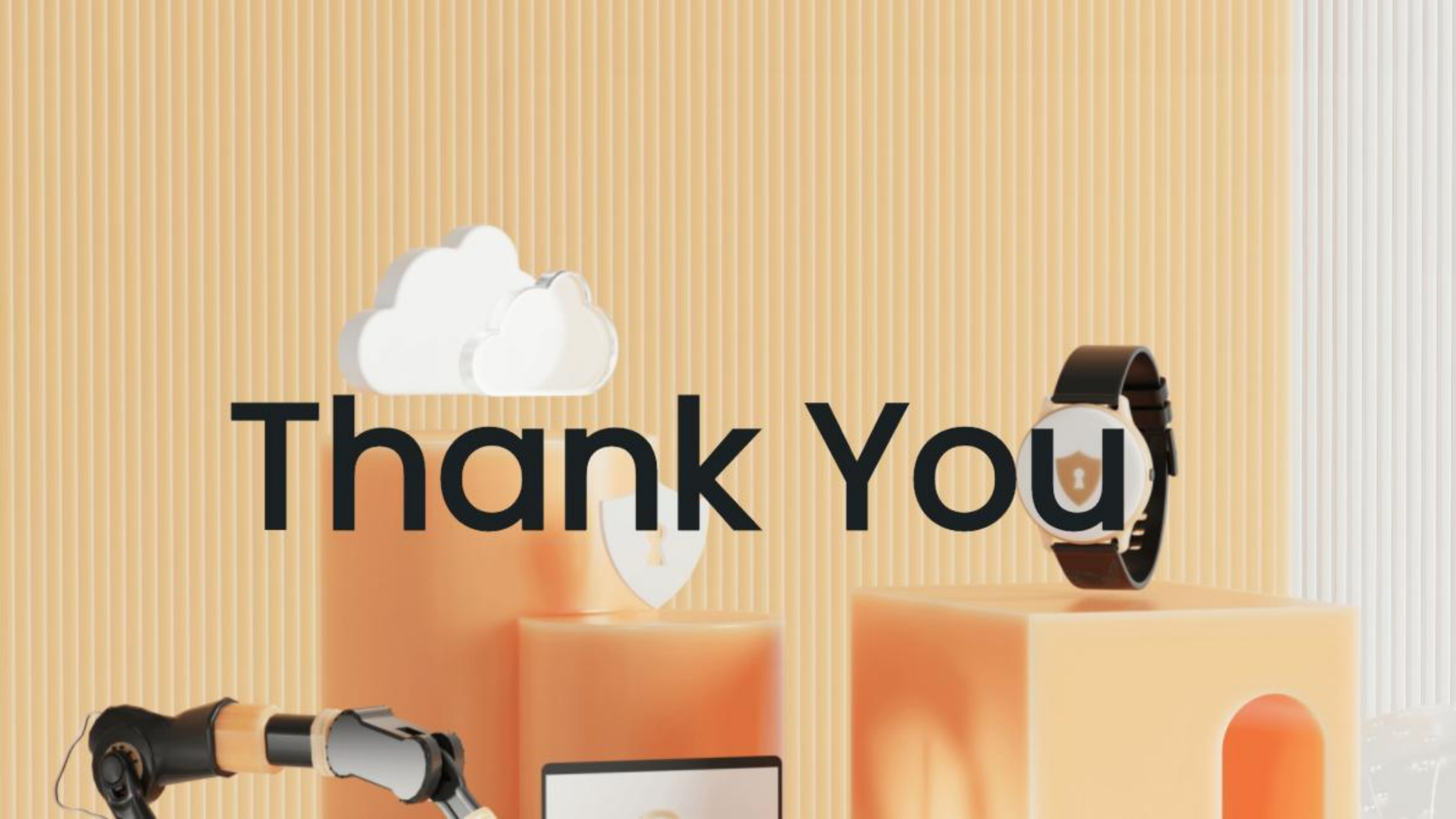
[1] R. Yadav and S. Kumari Sheoran, "Modified ARIMAModel for Improving Certainty in Spatio-TemporalCrime Event Prediction," 2018 3rd InternationalConference and Workshops on Recent Advances andInnovations in Engineering (ICRAIE), Jaipur, India,2018, pp. 1-4, doi: 10.1109/ICRAIE.2018.8710398.

[2] Dong, R. Ye and Y. Fan, "Impact of SpatialCorrelation on Crime Prediction in Communities withDifferent Crime Densities," 2022 IEEE 8th InternationalConference on Computer and Communications (ICCC),Chengdu, China, 2022, pp. 2227-2233, doi:10.1109/ICCC56324.2022.10065938.

[3]A. Sudhakar, D. C. Nandini, P. G. Bhavani and L. L.Priyanka, "Hybrid Crime Prediction Using GRU andARIMAX Models," 2025 8th International Conferenceon Trends in Electronics and Informatics (ICOEI),Tirunelveli, India, 2025, pp. 1510-1516, doi:10.1109/ICOEI65986.2025.11013209.

[4] S. Yao et al., "Prediction of Crime Hotspots based onSpatial Factors of Random Forest," 2020 15thInternational Conference on Computer Science &Education (ICCSE), Delft, Netherlands, 2020, pp. 811-815, doi: 10.1109/ICCSE49874.2020.9201899.

[5] K. T. M, L. T. N, M. Ithihas, N. R. Shetty, A. H. Nand S. Hebbar, "Crime Type and Occurrence PredictionUsing Machine Learning," 2024 Second InternationalConference on Advances in Information Technology(ICAIT), Chikkamagaluru, Karnataka, India, 2024, pp.1-5, doi: 10.1109/ICAIT61638.2024.10690652


Thank You