AMERICAN JOURNAL OF AI CYBER COMPUTING MANAGEMENT
Peer Reviewed, Rferred & Indexed Journal
E-ISSN:3069-0102
VOL.6, NO. 2(2026)

Ensemble Text-Visual Anomaly Detection for Authentic Real Estate Listings

P. Prasanna[1], K. Srinivas[2*], Ande Sathvika[3], Siddirala Sripriya[3], Bhukya Akash[3], Gantasala Adarsh[3]

[1]Assistant Professor, [2]Associate Professor & Head, [3]UG Student, [1,2,3]Department of Computer Science and Engineering (Data science)

[1,2,3]Vaagdevi Engineering College, Bollikunta, Warangal, 506005, Telangana, India

*Correspondence: K. Srinivas (srinivas_k@vecw.edu.in)

ABSTRACT

The rapid expansion of digital platforms in the real estate sector has intensified concerns regarding the authenticity of online property listings. Fraudulent advertisements not only result in financial losses for users but also erode trust in property platforms. Traditional verification methods, which rely on manual moderation or simple rule-based techniques such as keyword filtering, price comparison, and duplicate detection, are often inefficient and incapable of identifying complex fraud patterns. As the volume of listings grows, these conventional approaches struggle to ensure data reliability and platform security. To address these limitations, this study proposes an intelligent hybrid framework called ExtraNet for automated classification of real estate listings. The model integrates an Extra Trees Classifier (ETC) with an Artificial Neural Network (ANN) to leverage the strengths of both ensemble and deep learning techniques. The system processes structured tabular data containing key property attributes, including price, location, area size, number of rooms, amenities, and listing descriptions. A comprehensive preprocessing pipeline is applied to handle missing values, encode categorical variables, and normalize numerical features, ensuring improved model performance. For comparative analysis, baseline models such as Logistic Regression (LR), Nearest Centroid (NC) and K-Nearest Neighbors (KNN) are also implemented. The proposed solution is deployed as a user-friendly desktop application using a Tkinter-based graphical interface with (My Stretched Query Language) MySQL database integration for efficient data handling. Experimental results demonstrate that the ExtraNet model achieves a high classification accuracy of 99.61%, significantly outperforming traditional models. This approach provides a scalable and effective solution for detecting fraudulent listings and enhancing trust in digital real estate platforms.

Key words: Real Estate Fraud Detection, ExtraNet Model, Machine Learning (ML), Deep Learning (DL), Property Listing Verification, Tabular Data Classification.

1. INTRODUCTION

Online real estate platforms have become a vital resource for individuals seeking to buy or rent properties, significantly influencing customer behaviour and decision-making throughout the property search process. By 2023, the real estate market in Turkey reached an estimated value of USD 46 billion, reflecting substantial growth driven in part by increasing digital adoption, as shown in fig 1 The integration of digital technologies and online platforms has contributed to an annual rise of approximately 15% in real estate transactions. Notably, nearly one-third of all property sales in Turkey are now conducted through online channels, demonstrating the growing reliance on digital solutions within the industry.

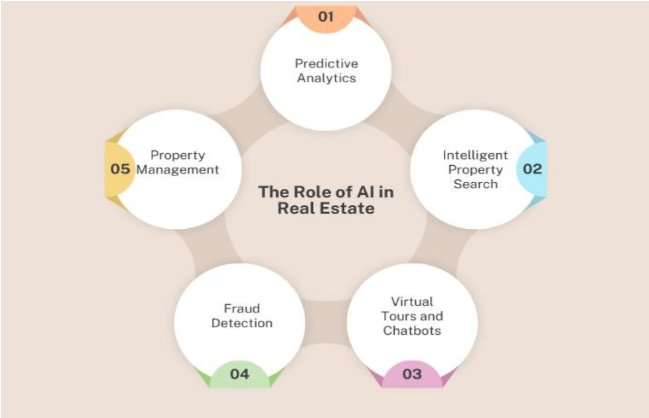


Fig. 1: AI in the real estate industry.

Compared to traditional methods, properties are promoted far more extensively on online platforms, with digital listings appearing at a much higher frequency and covering a broader segment of the market. Among the various elements of these listings, visual content such as images and virtual tours plays a crucial role in attracting potential buyers and tenants. In many cases, the quality and presentation of these visuals strongly influence a user's perception of a property and can determine whether they choose to explore it further. Furthermore, effective image classification and quality management are essential for enhancing user engagement and streamlining the listing process. Well-structured and visually appealing property images not only capture attention but also improve browsing efficiency and reduce the time required for manual editing and review. High-quality visuals enable users to better evaluate properties, leading to more informed decisions. As a result, real estate agencies must prioritize the optimization and organization of images to improve user experience and maximize the effectiveness of their online listings.

2. LITERATURE SURVEY

Küp, et al. [1] proposed an extended platform that examined the role of machine learning models and image enhancement methods in real estate technology. The authors also introduced an alternative solution that could be integrated into intelligent listing systems to enhance user experience and information accuracy. Their study demonstrated that artificial intelligence and machine learning could be effectively combined for cloud-distributed services, thereby paving the way for future innovations in the real estate sector and intelligent marketplace platforms. Cristian Bogdan et al., [2] aimed to identify how fake news appeared in a world where access to information was simple and individuals could easily verify the truthfulness of information. The study also examined the barriers used to protect businesses from being affected and customers from being misinformed. Data for the study were collected through in-depth interviews with entrepreneurs from real estate businesses. The findings revealed that the increasing level of fake news negatively affected the trust-building process. The authors recommended that organizations analyze user behavior more thoroughly and utilize it in the decision-making process to reduce the negative effects of fake news. They also suggested that customers learn to identify quality brand communication, pay attention to details, and exercise caution when making decisions based on potentially fake news. Zhang X, et al. [3] proposed and evaluated four image–text similarity measures, namely textual similarity, semantic similarity, contextual similarity, and post-training similarity. The textual and semantic similarities represented the original image text relationships in terms of textual content and image caption information, while the contextual similarity reflected the image–text similarity based on meaningful named entities. Consumers with higher ecological motivation and environmental awareness are more prone to engage in green purchase behavior. A study on green food purchase found that pro-environmental self-identity predicts green purchase intention [4]. The intention of the consumer to buy eco-friendly products is to support environmental sustainability, thus impacting pro-environmental self-identity. So, pro-environmental purchasing is more inclined to green purchase decisions [5]. The key to self-identity is perceiving oneself as possessing specific traits: characteristics, qualities, and abilities in combination with personality traits, physical attributes interests, hobbies, memories, moral qualities, social factors, and cognitive processes that are selectively chosen to identify themselves [6].

Karpook et al. [7] Investment in AI and machine learning (ML) was having a transformative effect on corporate and commercial real estate, providing unprecedented efficiencies in tasks such as market analysis, property valuation, document automation, and lease abstraction. These activities, which had always been tedious and time-consuming manual processes, could be done with great precision and speed using the emerging technologies, freeing workers for tasks that relied on human interaction and judgment. These innovations, which were overwhelmingly data-driven, were particularly well-suited to real estate, with its heavy reliance on property and market data. The promise of AI and ML was driving significant investment in an industry that had long stood out for its resistance to new technologies, with predictions of investment exceeding US$700bn by 2028. While exciting, this new territory was not without danger; the emergence of 'deep fakes' that used the same AI and ML technologies to enable fraud and theft threatened businesses that were just beginning to understand how to protect themselves. Joel Paul et al. [8] successful execution of financial and real estate machine learning (ML) prediction models required perfect performance balance to produce accurate results and reliable interpretations. The applications of financial ML models spanned algorithmic trading combined with credit risk evaluation as well as fraud detection, and real estate ML models concentrated on property evaluation, market trend analysis, and rental price forecasting. The existing models encountered obstacles because they experienced overfitting and encountered problems in data quality and had to follow regulatory guidelines. The author examined multiple techniques which enhanced model efficiency and explained the applications of feature engineering together with regularization combined with robust validation methods along with hybrid modeling approaches. Patel, et al. [9] aimed to explore whether signs of property market bubble formation had become evident, and to create a pragmatic predictive model of Irish real-estate valuation. Acknowledging the challenge of the market, this study represented a holistic approach aware of the many limitations of traditional investigation that created a one-sided distinction between the factors influencing property values.

F. A. Sulistio et al. [10] proposed a multimodal deep learning model that combined textual features from Indonesia Bidirectional Encoder Representations from Transformers (IndoBERT) with visual features extracted from a Residual Network-18 (ResNet18) based Convolutional Neural Network (CNN) to classify listings as real or fake. The dataset comprised 2,632 entries, consisting of 1,632 real listings and 1,000 synthetically generated fake ones, and yielded a classification accuracy of 97.99%. This result outperformed unimodal baselines that used only text (97.15%) or image (90.51%) inputs. Duy Nguyen et al. [11] proposed FADAML, a novel end-to-end machine learning system to detect and filter out fake online advertisements. Their system combined techniques in multimodal machine learning and automated machine learning to achieve a high detection rate. As a case study, the authors applied FADAML to detect fake advertisements on popular Vietnamese real estate websites. Their experiments showed that the system achieved 91.5% detection accuracy, which significantly outperformed three different state-of-the-art fake news detection systems. Yuan et al. compared different approaches using regression and classification models for property price prediction [12]. determined the most effective method by evaluating different machine learning techniques. The result proved the usability and the performance of machine learning models for the real estate industry in forecasting property prices. Lemeš and Akagic predicted the cost of proprieties

AMERICAN JOURNAL OF AI CYBER COMPUTING MANAGEMENT
Peer Reviewed, Rferred & Indexed Journal
E-ISSN:3069-0102
VOL.6, NO. 2(2026)

and assessed the performances of the algorithms. The prediction accuracy was increased by centering on feature engineering and model optimization [13]. Law et al. built prediction models for housing prices using deep learning models, street and satellite images, and economic factors [14]. According to the previous results, exterior and top-view real estate images are used to effectively evaluate properties. Semnani and Rezaei studied real estate price prediction using satellite images and deep learning techniques [15]

3. PROPOSED SYSTEM

The proposed methodology presents a structured and data-driven framework for classifying real estate listings as genuine or fraudulent using machine learning techniques. The framework follows a systematic pipeline that begins with dataset acquisition and progresses through preprocessing, feature engineering, model training, and prediction stages. The overall architecture of the system is illustrated in Fig. 2. By integrating multiple machine learning models with effective data preprocessing strategies, the methodology enables accurate analysis of heterogeneous real estate data. The framework is designed to handle missing values, mixed data types, and large-scale datasets commonly observed in real-world property listings. Comparative evaluation across multiple models ensures reliability and consistency in classification performance. The system supports both administrative operations for model training and user-level prediction, making it suitable for practical deployment in real-time environments.

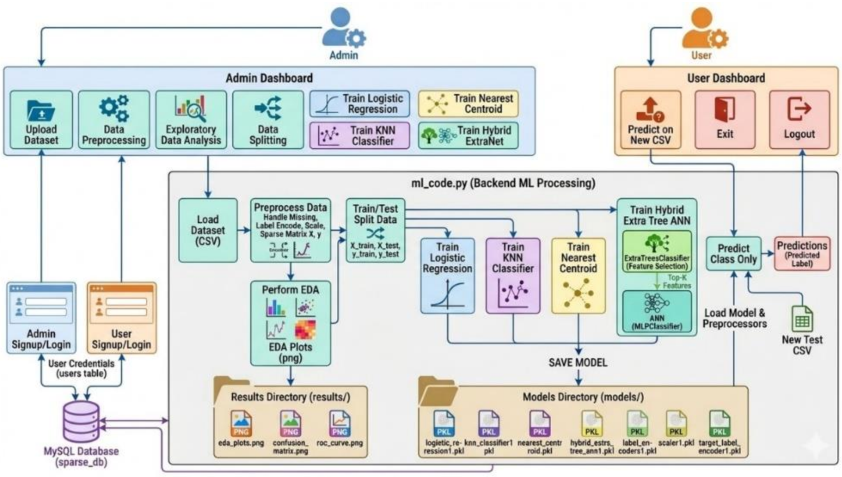


Fig. 2: Proposed System Architecture.

User Interface (Admin/User Interaction)

- Users interact with the system through a graphical interface supporting both Admin and User roles.
- **Admin users** are granted elevated permissions to upload master datasets, perform preprocessing, conduct Exploratory Data Analysis (EDA), and trigger model training.
- **General users** can upload specific listing data to obtain real-time classification predictions.
- All user inputs are captured via the GUI and forwarded to the backend for computational analysis.

Backend Processing Environment

- This environment serves as the orchestration layer, managing the complete workflow from data ingestion to the final prediction output.
- It coordinates the sequence of preprocessing, feature transformation, model training, and inference.

- The backend handles model serialization (loading and saving .pkl or .h5 files) and ensures efficient communication between the frontend and the machine learning modules.

Raw Dataset Input

- The primary data source consists of real estate listing data containing a mix of numerical and categorical features.
- Attributes include price, location details, number of rooms, and specific listing descriptions.
- The raw data is audited for missing values, inconsistencies, and noise before being passed into the analytical pipeline.

Data Preprocessing Module

- This module prepares the data for mathematical modeling by resolving inconsistencies.
- Missing values are addressed through strategic removal or imputation techniques.
- Categorical variables are transformed into numerical format using **Label Encoding**, while numerical features are standardized using scaling methods to ensure feature parity.
- The output is a clean, structured dataset optimized for high-performance training.

Feature Engineering and Representation

- The system separates categorical and numerical features to apply appropriate transformations to each type.
- Transformed features are combined using sparse matrix representation, which allows for the efficient handling of high-dimensional real estate data.
- These processed feature vectors serve as the direct input for the classification algorithms.

Data Splitting and Preparation

- To ensure an unbiased evaluation, the dataset is divided into training and testing subsets.
- This split maintains a consistent data distribution, ensuring that the patterns learned during training are accurately validated against unseen test data.

Baseline Classification Models

- The framework implements several standard machine learning models to establish performance benchmarks:
    - **LR:** For linear decision boundaries.
    - **KNN:** For distance-based similarity classification.
    - **NC:** For centroid-based pattern matching.
- Each model independently processes the training data to generate comparative predictions.

Hybrid Model (ExtraNet: ETC + ANN)

The core of the system is the ExtraNet hybrid model, which operates in two distinct phases:

1. **Feature Selection (ETC):** The Extra Trees Classifier is used to estimate feature importance and select only the most relevant attributes, reducing dimensionality.
2. **Deep Learning Classification (ANN):** An Artificial Neural Network is applied to these selected features to capture complex, non-linear relationships, significantly enhancing prediction accuracy.

Prediction Output and Classification

- The system assigns a final label to each property listing: Fake or Real.

- The architecture supports both single-instance predictions and large-scale batch processing.
- Results are displayed immediately on the graphical interface and can be stored for long-term audit and analysis.

Model Evaluation and Performance Analysis

- The framework rigorously assesses all models using metrics such as Accuracy, Precision, Recall, and F1-score.
- Confusion matrices and ROC curves are generated to provide a visual representation of the system's ability to distinguish between real and fraudulent listings.
- This comparative analysis ensures that the most reliable model is utilized for the live production environment.

4.1 Hybrid ExtraNet Model

The hybrid ExtraNet Model integrates the strengths of ETC and an ANN to achieve superior detection accuracy. The ETC algorithm captures complex decision boundaries and reduces variance through ensemble averaging, while the ANN layer enhances non-linear feature learning and pattern recognition. This hybrid approach ensures improved robustness, adaptability, and generalization. The ETC first performs feature selection and builds multiple decision trees using random splits for attributes. It outputs probabilistic feature importance values and predictions, as shown in fig. 3. These predictions are then fed into an ANN that learns higher-order, non-linear representations of the data. The ANN's dense layers use activation functions such as ReLU and sigmoid to refine classification boundaries and output final predictions.

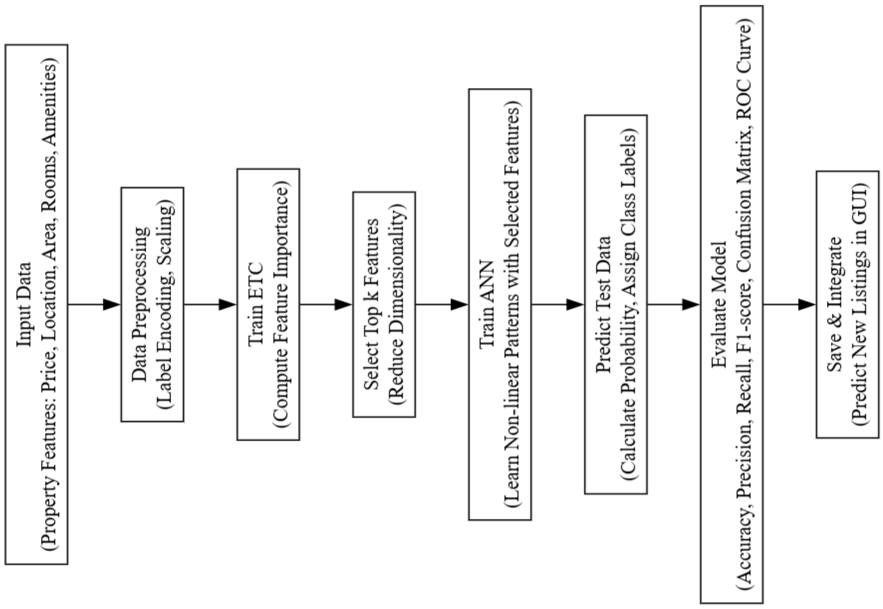


Fig. 3: Internal workflow of hybrid ExtraNet model.

Input Feature Selection: The pre-processed dataset containing numerical and categorical property features such as price, location, area, number of rooms, and amenities is used as input. Categorical features are encoded, and numerical features are standardized to ensure consistency.

Feature Importance Extraction using ETC: The ETC is trained on the input dataset to compute feature importance scores. This step identifies which property attributes contribute most to distinguishing real and fake listings.

AMERICAN JOURNAL OF AI CYBER COMPUTING MANAGEMENT
Peer Reviewed, Rferred & Indexed Journal
E-ISSN:3069-0102
VOL.6, NO. 2(2026)

Feature Selection: Based on the feature importance scores from ETC, the top k most relevant features are selected. This reduces dimensionality, eliminates noise, and focuses the subsequent ANN on the most informative features.

Training the ANN: The selected top features are fed into an ANN model. The ANN consists of multiple layers and neurons that learn non-linear patterns and complex relationships between property features and listing authenticity. The model is trained iteratively using backpropagation and optimization techniques.

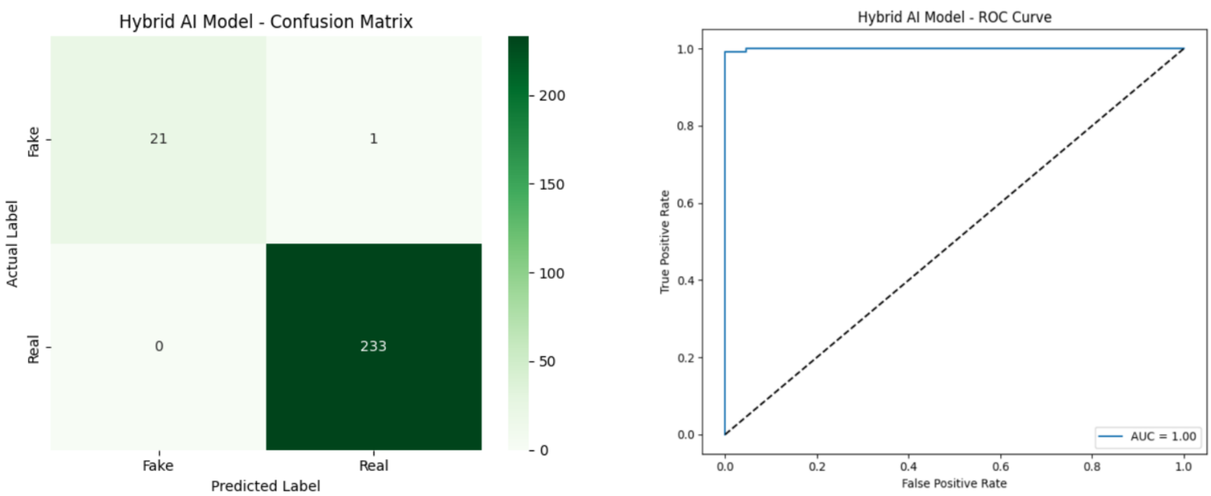
Prediction on Test Data: Once trained, the Hybrid ExtraNet model predicts the authenticity of listings in the test dataset. The ANN outputs probabilities, which are converted into class labels (real or fake) based on a threshold.

Model Evaluation: The predicted labels are compared with actual labels, and performance metrics such as accuracy, precision, recall, and F1-score are calculated. Confusion matrices and ROC curves are also generated to assess model effectiveness.

Integration into the System: The trained Hybrid ExtraNet model is saved and integrated into the GUI system. It is used to predict the authenticity of new property listings uploaded by users, providing a robust and accurate method for fraud detection.

4. RESULTS AND DISCUSSION

Fig. 4 the confusion matrix highlights hybrid ExtraNet predictions, showing counts of real and fake listings classified correctly and incorrectly. True positives and true negatives confirm high accuracy, while false positives and false negatives indicate rare misclassifications. The ROC curve shows sensitivity versus specificity, and the AUC demonstrates the superior performance of hybrid ExtraNet, combining ETC feature selection and ANN non-linear pattern recognition for accurate classification.



(a) **(b)**

Fig. 4: Obtained confusion matrix and ROC curves from hybrid ExtraNet model.

Predictions:

	ID	Listing Confirmation Method	Deposit Amount	...	Listing Weekday	Days Since Listed	Predicted Label
0	TRAIN_0004	On-site Confirmation	346000000	...	Tuesday	382	Real
1	TRAIN_0012	Phone Confirmation	348500000	...	Monday	362	Real
2	TRAIN_0022	On-site Confirmation	24500000	...	Sunday	426	Fake
3	TRAIN_0031	On-site Confirmation	132000000	...	Wednesday	367	Fake
4	TRAIN_0047	On-site Confirmation	76000000	...	Tuesday	571	Fake
5	TRAIN_0048	On-site Confirmation	156000000	...	Friday	512	Fake
6	TRAIN_0093	Phone Confirmation	403500000	...	Wednesday	437	Real
7	TRAIN_0106	On-site Confirmation	12500000	...	Saturday	469	Fake
8	TRAIN_0117	On-site Confirmation	12000000	...	Sunday	419	Fake
9	TRAIN_0118	On-site Confirmation	11000000	...	Friday	449	Fake
10	TRAIN_0128	서류확인	69500000	...	Wednesday	360	Fake
11	TRAIN_0135	On-site Confirmation	406500000	...	Thursday	478	Real
12	TRAIN_0136	On-site Confirmation	135000000	...	Friday	449	Fake
13	TRAIN_0147	On-site Confirmation	154500000	...	Friday	337	Fake
14	TRAIN_0156	On-site Confirmation	8000000	...	Friday	736	Fake
15	TRAIN_0160	Phone Confirmation	11000000	...	Sunday	293	Fake
16	TRAIN_0164	On-site Confirmation	26000000	...	Saturday	406	Fake
17	TRAIN_0000	On-site Confirmation	402500000	...	Wednesday	276	Real

Fig. 5: Sample predictions on test data.

Fig. 5 showcases the final output of the system. It displays the dataset entries along with their predicted labels Real or Fake based on the model's analysis. The results are presented in a tabular format for easy interpretation, with additional options for saving or exporting the prediction report. This screen reflects the system's ability to process unseen data and deliver reliable classification outcomes instantly. It validates the overall research objective ensuring authenticity and transparency in online real estate listings through an intelligent AI-driven solution.

Table 1: Performance comparison for the LR, NC, KNN, and hybrid ExtraNet models.

Algorithm Name	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
LR Model	8.63	4.31	50.00	7.94
KNN Model	91.37	45.69	50.00	47.75
NC Model	77.25	61.41	79.32	62.18
Hybrid ExtraNet	99.61	99.79	97.73	98.73

Table 1 presents a detailed performance comparison of the four models implemented in the research LR, NC, KNN, and hybrid ExtraNet. LR showed limited performance, achieving only 8.63% accuracy with low precision and F1-score, indicating its inability to capture complex, non-linear patterns in the real estate data. NC achieved 77.25% accuracy with 61.41% precision and 62.18% F1-score, demonstrating strong detection of real listings while showing lower precision for fake listings. KNN performed significantly better, achieving 91.37% accuracy, demonstrating that neighborhood-based learning effectively identified similarities among authentic listings. However, moderate precision and recall reflect its limitations in consistently classifying boundary cases. The hybrid ExtraNet model outperformed all others, achieving 99.61% accuracy, 99.79% precision, and 98.73% F1-score. The combination of ETC feature selection and ANN learning enabled the model to generalize effectively, distinguishing real listings from fraudulent ones with near-perfect reliability.

REFERENCES

[1] Küp, E.T.; Sözdinler, M.; Işık, A.H.; Doksanbir, Y.; Akpınar, G. Enhancing Real Estate Listings Through Image Classification and Enhancement: A Comparative Study. *Eng. Proc.* **2025**, *92*, 80. https://doi.org/10.3390/engproc2025092080

[2] Cristian Bogdan ONETE The Department of Business, Consumer Sciences and Quality Management, Faculty of Business and Tourism, Bucharest University of Economic Studies, Bucharest, Romania cristian.onete@com.ase.ro

[3] Zhang X, Dadkhah S, Weismann AG, Kanaani MA, Ghorbani AA. 2024. Multimodal fake news analysis based on image-text similarity. IEEE Transactions on Computational Social Systems 11(1):959-972

[4] Bajar, R.G.; Ong, A.K.; German, J.D. Determining sustainable purchase behavior for green products from name-brand shops: A Gen Z perspective in a developing country. Sustainability 2024, 16, 3747.

[5] Lopes, J.M.; Gomes, S.; Trancoso, T. Navigating the green maze: Insights for businesses on consumer decision-making and the mediating role of their environmental concerns. Sustain. Account. Manag. Policy J. 2024, 15, 861–883.

[6] Rudnev, M.; Barrett, H.C.; Buckwalter, W.; Machery, E.; Stich, S.; Barr, K.; Bencherifa, A.; Clancy, R.F.; Crone, D.L.; Deguchi, Y.; et al. Dimensions of wisdom perception across twelve countries on five continents. *Nat. Commun.* **2024**, *15*, 6375.

[7] Karpook, Corporate Real Estate Journal, Volume 14 / Number 3 / Spring 2025, pp. 210-221(12), Henry Stewart Publications DOI**:** https://doi.org/10.69554/TQFL5101

[8] Joel Paul Balancing Performance in Financial and Real Estate Machine Learning Models, February 2025.

[9] Patel, Diepesh Manish (2024) Stacking Ensemble Approach towards Predicting Irish-Real Estate. Masters thesis, Dublin, National College of Ireland.

[10] F. A. Sulistio, S. Suakanto, N. Ambarsari and S. C. R. Rijadi, "Detecting Fake Boarding House Listings Using Multimodal Deep Learning," 2025 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology *(IAICT)*, Bali, Indonesia, 2025, pp. 330-337, doi: 10.1109/IAICT65714.2025.11101391.

[11] Duy Nguyen, Fake Advertisements Detection Using Automated Multimodal Learning: A Case Study for Vietnamese Real Estate Data, Datal, Ho Chi Minh City, Vietnam duyndv@datal.com.vn.

[12] Yuan, F.; Wu, J.; Wei, Y.D.; Wang, L. Policy change, amenity, and spatiotemporal dynamics of housing prices in Nanjing, China. Land Use Policy, 75, 225–236.

[13] Lemeš, L.; Akagic, A. Prediction of Real Estate Market Prices with Regression Algorithms; Lecture Notes in Networks and Systems; Springer: Cham, Switherland, 2023; Volume 539, pp. 401–411.

[14] Law, S.; Paige, B.; Russell, C. Take a look around: Using street view and satellite images to estimate house prices. ACM Trans. Intell. Syst. Technol. 2019, 10, 54.

[15] Semnani, S.J.; Rezaei, H. House Price Prediction using Satellite Imagery. arXiv 2021, arXiv:2105.06060.

[16] Liu, Y.; Gross, L.; Li, Z.; Li, X.; Fan, X.; Qi, W. Automatic Building Extraction on High-Resolution Remote Sensing Imagery Using Deep Convolutional Encoder-Decoder with Spatial Pyramid Pooling. IEEE Access, 7, 128774–128786.

[17] Zhu, X.X.; Tuia, D.; Mou, L.; Xia, G.S.; Zhang, L.; Xu, F.; Fraundorfer, F. Deep learning in remote sensing: A review. IEEE Geosci. Remote Sens. Mag, 5, 8–36.