MedSureAI: A Hybrid Ensemble Machine Learning Framework with Anomaly Detection for Real-Time Health Insurance Fraud Detection

[1] Dr. M.K. Jayanti Kannan, [2]Akanksha Dhote
[1] *Professor* [2] *PG Student School of Computing Science and Engineering and Artificial Intelligence, VIT Bhopal University, Bhopal-Indore Highway, Kothri Kalan, Sehore, Madhya Pradesh – 466114*

***Abstract*— Health insurance fraud has emerged as a critical challenge in modern healthcare systems, resulting in significant financial losses, increased operational costs, and reduced trust among stakeholders. Traditional fraud detection methods, primarily based on rule-based systems and manual auditing, are inadequate for identifying complex and evolving fraud patterns in large-scale healthcare datasets. This research proposes an advanced Artificial Intelligence (AI) driven framework, MedSureAI, designed to enable accurate and real-time fraud detection in health insurance systems. The proposed approach integrates multiple stages, including data preprocessing, advanced feature engineering, hybrid class imbalance handling, and ensemble machine learning techniques. Statistical, behavioral, and risk-based features are extracted to capture hidden fraud patterns. A hybrid imbalance handling strategy combining SMOTE and undersampling is employed to address the skewed distribution of fraudulent and non-fraudulent cases. The model architecture incorporates a combination of Random Forest, XGBoost, and Isolation Forest algorithms, enabling the detection of both known and unknown fraud patterns through supervised and unsupervised learning. The ensemble model demonstrates superior performance compared to individual models, achieving high accuracy, improved recall, and strong ROC-AUC scores. The system is further deployed using a Streamlit-based interface, enabling real-time fraud prediction and automated decision support. Experimental results validate the effectiveness, scalability, and practical applicability of the proposed framework. The study concludes that the integration of hybrid machine learning models with real-time deployment provides a robust and efficient solution for healthcare fraud detection, with potential for further enhancement using advanced AI techniques.**

***Keywords*— Health Insurance Fraud Detection, Machine Learning, Artificial Intelligence, Ensemble Learning, Random Forest, XGBoost, Isolation Forest, SMOTE, Class Imbalance Handling, Feature Engineering, Anomaly Detection, Real-Time Prediction, Streamlit Deployment, Healthcare Analytics, Predictive Modeling.**

I. INTRODUCTION

The exponential growth of digital healthcare ecosystems has significantly transformed the way medical services are delivered, recorded, and reimbursed. Modern health insurance systems generate vast volumes of heterogeneous data, including patient demographics, clinical records, billing transactions, and provider-level activities. While this digital transformation has improved operational efficiency and accessibility, it has simultaneously introduced critical vulnerabilities most notably, health insurance fraud. Health insurance fraud is a pervasive issue that results in billions of dollars in financial losses annually across global healthcare systems. Fraudulent activities may include exaggerated billing, phantom claims, unnecessary medical procedures, identity theft, and collusion between providers and beneficiaries. These fraudulent behaviors not only impose financial burdens on insurance companies but also degrade trust in healthcare systems and lead to increased premiums for policyholders. Traditional fraud detection mechanisms, primarily based on rule-based systems and manual auditing, are no longer sufficient to handle the scale, complexity, and evolving nature of fraud patterns. These systems are inherently reactive, limited to predefined rules, and incapable of adapting to new fraud strategies. Moreover, the increasing volume of claims requires real-time decision-making, which is beyond the capability of manual verification processes.

To address these challenges, this research proposes an Artificial Intelligence (AI)-driven framework, termed MedSureAI, designed to enable real-time fraud detection and autonomous decision support in health insurance systems. By leveraging advanced machine learning techniques such as ensemble learning, feature engineering, and class imbalance

handling, the system aims to detect hidden fraud patterns and provide actionable insights during claim processing.

II. LITERATURE REVIEW OF EXISTING SYSTEMS

Title of the Paper	Problem Statement	Objective	Technology Used	Methodology Used	Efficiency / Performance	Issues / Limitations
Predicting Insolvency of Insurance Companies in Egyptian Market Using Bagging and Boosting Ensemble Techniques	Insurance companies face significant risks of financial insolvency, which impacts stakeholders and economic stability. Existing models either lack accuracy, do not use ensemble	- Develop a predictive model for insurance insolvency in Egypt. - Introduce the first public dataset for Egyptian insurance companies. - Compare classical ML and ensemble learning methods. - Identify key financial ratios influencing insolvency	- Programming: Python - Libraries: Scikit-learn, Pandas, Matplotlib, CatBoost - Models: SVM, Random Forest, Bagging, Boosting (CatBoost)	- Data collection (1999–2019, 11 companies, 22 ratios) - Data preprocessing (cleaning, encoding) - Feature correlation analysis - Model training using ML & ensemble methods - Hyperparameter tuning using Grid Search - Validation: Hold-out & K-fold cross-validation	- Ensemble models outperform classical ML models - CatBoost achieved best performance - Metrics used: MAE, RMSE, R^2 - High predictive accuracy with low error rates	- Dataset limited to Egyptian market only - Imbalanced data across companies - Limited number of samples - Some models (e.g., SVM) performed poorly due to data imbalance

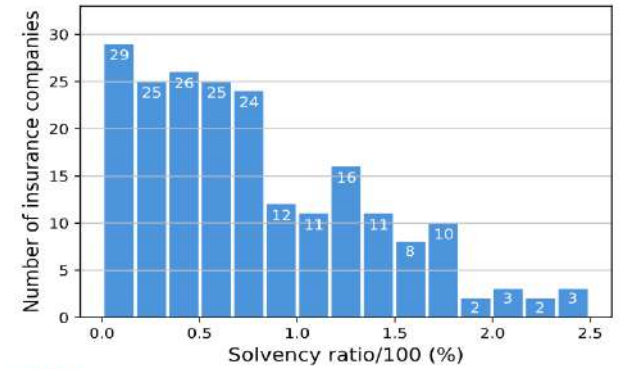


FIGURE 1. Histogram chart of the values of solvency ratio in the dataset.

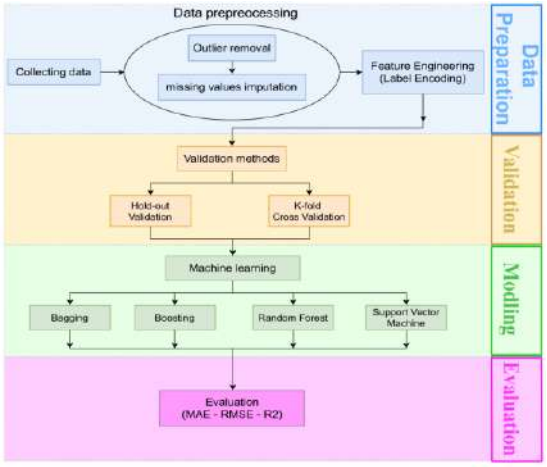
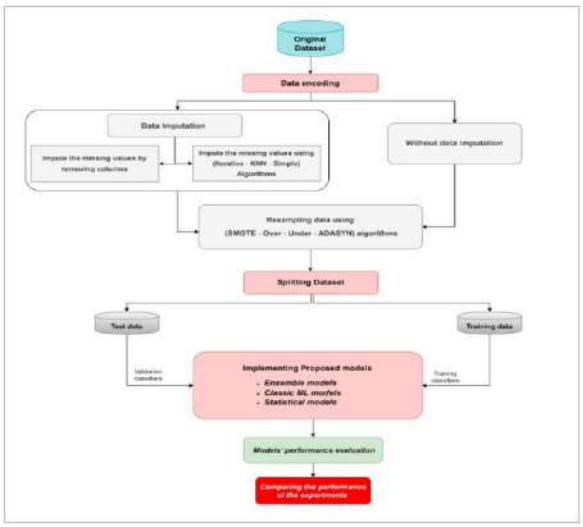


FIGURE 3. Block diagram of the proposed method.

Fig.1: Predicting Insolvency of Insurance Companies in Egyptian Market Using Bagging and Boosting Ensemble Techniques

Title of the Paper	Problem Statement	Objective	Technology Used	Methodology Used	Efficiency / Performance	Issues / Challenges
Machine Learning Based Method for Insurance Fraud Detection on Class Imbalance Datasets With Missing Values	Insurance fraud detection is challenging due to class imbalance (few fraud cases) and missing data, which reduce model accuracy and reliability. Existing approaches do not adequately handle both issues together or analyze overfitting.	- Develop an effective fraud detection system using ML. - Handle class imbalance and missing data problems. - Compare different imputation and resampling techniques. - Perform overfitting analysis. - Improve prediction accuracy over existing methods.	- Programming: Python - Libraries: Scikit-learn, Pandas - Models: Random Forest, SVM, KNN, Logistic Regression, Ridge Regression - Ensemble Methods: Bagging, Boosting, Stacking - Techniques: SMOTE, ADASYN, Random Sampling	- Data collection (real dataset + Kaggle dataset) - Data preprocessing (cleaning, encoding) - Handling missing values (imputation + column removal) - Handling class imbalance (SMOTE, oversampling, undersampling, ADASYN) - Feature importance & correlation analysis - Model training (ML + ensemble models) - Validation: Hold-out & K-fold cross-validation - Evaluation using multiple metrics and overfitting analysis	- Handling class imbalance significantly improves model performance. - Missing value handling has moderate impact. - Proposed models outperform existing methods. - Ensemble models achieve higher accuracy. - Improved fraud detection with better precision and recall.	- Highly imbalanced datasets (fraud cases very low). - Presence of missing and noisy data. - Risk of overfitting in ML models. - Limited availability of real-world datasets. - High dependency on preprocessing techniques.



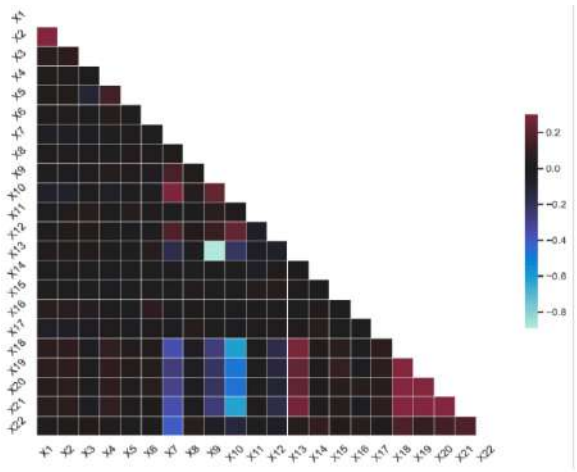


FIGURE 2. Correlation matrix of Dataset-1 features.

Title of the Paper	Problem Statement	Objective	Technology Used	Methodology Used	Efficiency / Performance	Issues / Challenges
Cost-Sensitive Learning for Medical Insurance Fraud Detection With Temporal Information	Healthcare fraud in systems like Medicare leads to billions of dollars in losses annually. Traditional fraud detection methods are inefficient, fail to utilize temporal data, and do not consider financial cost trade-offs in decision-making.	- Develop an effective fraud detection model using Medicare data. - Incorporate temporal information from past years. - Apply cost-sensitive learning to reflect real financial impact. - Improve feature extraction using advanced statistical techniques. - Reduce financial losses due to fraud.	- Programming: Python (implied) - Techniques: Functional Data Analysis - Algorithms: Machine Learning classifiers - Methods: Functional Principal Component Analysis (FPCA), Cost-Sensitive Learning - Dataset: Medicare claims data (CMS, Part D dataset)	- Data collection from Medicare Part D dataset (2013–2018) - Labeling fraud using LEIE database - Feature engineering using temporal trajectories of data - Application of FPCA for extracting temporal features - Distributional FPCA for probability density features - Handling class imbalance using random undersampling - Model training using multiple ML algorithms - Implementation of cost-sensitive learning framework using cost matrix - Evaluation based on predictive performance and cost savings	- Achieved reasonably good prediction performance. - Significant cost savings (~55%) using cost-sensitive approach. - Outperformed traditional non-cost-sensitive models. - Effective use of temporal data improved prediction quality. - Better decision-making by balancing fraud detection vs investigation cost.	- Extreme class imbalance (only ~0.025% fraud cases). - High complexity in handling temporal data. - Dependency on accurate estimation of cost parameters. - Large-scale healthcare data processing challenges. - Traditional models may fail without proper feature engineering.

Fig.2: Machine Learning Based Method for Insurance Fraud Detection on Class Imbalance Datasets With Missing Values

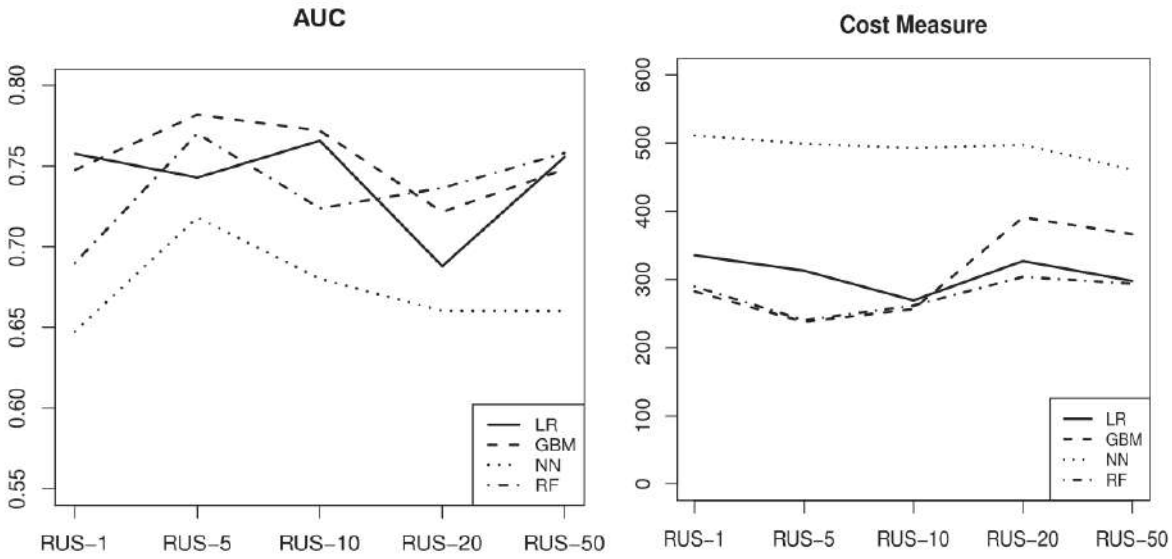


Fig.3: Cost-Sensitive Learning for Medical Insurance Fraud Detection With Temporal Information

Title of the Paper	Problem Statement	Objective	Technology Used	Methodology Used	Efficiency / Performance	Issues / Challenges
Going Digital: Case Study of an Italian Insurance Company	Traditional insurance companies have failed to innovate, especially in customer experience. With digital transformation and rising customer expectations, existing systems are outdated, complex, and not user-friendly. Companies struggle to adopt new technologies and respond to emerging risks like cyber threats.	- Identify digital transformation strategies for insurance companies. - Improve customer experience using digital solutions. - Analyze how small insurance firms can adopt InsurTech. - Develop a prototype platform to connect customers and company. - Expand services to include modern risks like cyber insurance.	- Digital technologies (Big Data, Cloud Computing, Social Media) - InsurTech solutions - Design Thinking & Innovation tools - Prototype development (customer interaction platform) - Cyber risk insurance technologies	- Case study of an Italian insurance company (Assinord Verona) - Analysis of company value chain - Interviews with customers (15 sessions) and employees - Brainstorming sessions for idea generation - Design thinking approach for innovation - Development and testing of prototype platform - Evaluation of strategies: internal innovation, buy, partner, invest in startups - Study and integration of cyber risk insurance products	- Improved customer engagement through digital platform. - Faster and more efficient insurance recommendation process. - Increased employee productivity and reduced dependency on experts. - Positive feedback from customers and employees on prototype. - Successful introduction of cyber insurance products with early adoption.	- Resistance to change within the organization. - Difficulty in adopting new technologies and mindset shift. - Limited internal capabilities for innovation. - Complexity in integrating digital solutions. - Need for continuous adaptation to evolving customer expectations.

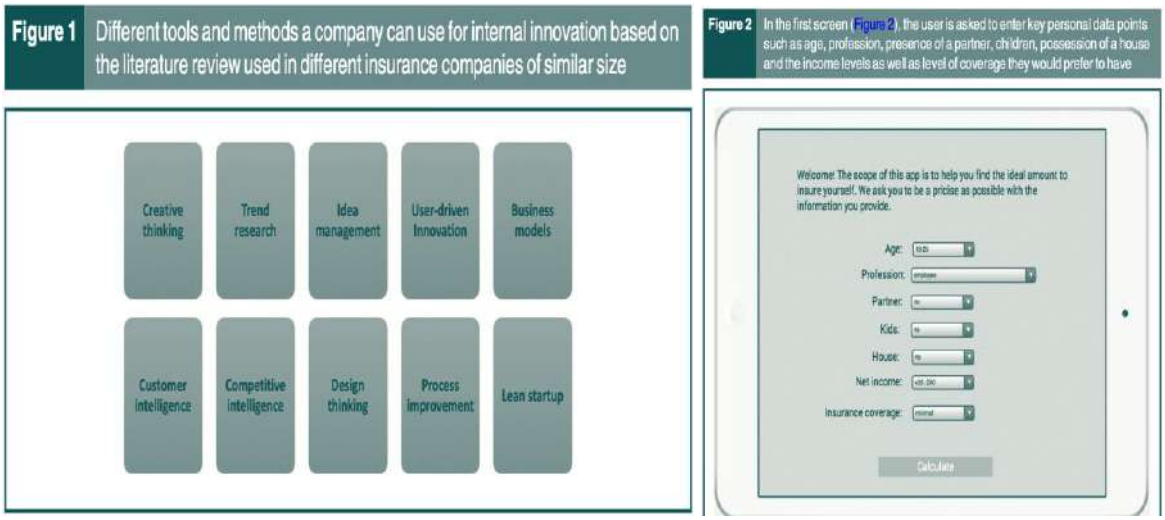


Fig.4: Going Digital: Case Study of an Italian Insurance Company

Title of the Paper	Problem Statement	Objective	Technology Used	Methodology Used	Efficiency / Performance	Issues / Challenges
Foundational AI in Insurance and Real Estate: A Survey of Applications, Challenges, and Future Directions	Traditional insurance and real estate systems are inefficient, manual, and lack intelligent decision-making. There is limited integration of AI for automation, risk assessment, fraud detection, and property valuation. Additionally, challenges like data inconsistency, lack of explainability, and regulatory constraints hinder AI adoption.	• To provide a comprehensive survey of AI applications in insurance and real estate. • To analyze technologies like ML, DL, NLP, and CV. • To identify challenges (technical, ethical, regulatory). • To propose future research directions and AI adoption roadmap.	• Machine Learning (Regression, Classification, Clustering) • Deep Learning (CNN, RNN, LSTM) • Natural Language Processing (NLP) • Computer Vision • Reinforcement Learning • IoT Integration (Smart Buildings) • Explainable AI (SHAP) • Ensemble Models (Random Forest, Gradient Boosting)	• Survey-based analytical study • Comparative analysis of AI techniques • Use-case driven evaluation (insurance + real estate) • Mathematical modeling (Bayesian inference, regression, optimization) • Case studies and industry examples • Identification of research gaps and future trends	• Improved risk assessment accuracy using ML models • Faster claims processing and underwriting automation • Enhanced fraud detection via predictive analytics • Accurate property valuation using regression & DL models • Cost reduction and operational efficiency through automation • Real-time decision-making and dynamic pricing improvements	• Poor data quality (incomplete, inconsistent datasets) • Lack of standardized datasets and benchmarks • Model interpretability (black-box AI problem) • Regulatory and compliance constraints • Data silos across departments • Ethical concerns (bias, fairness) • Difficulty in cross-market generalization • Integration with legacy systems

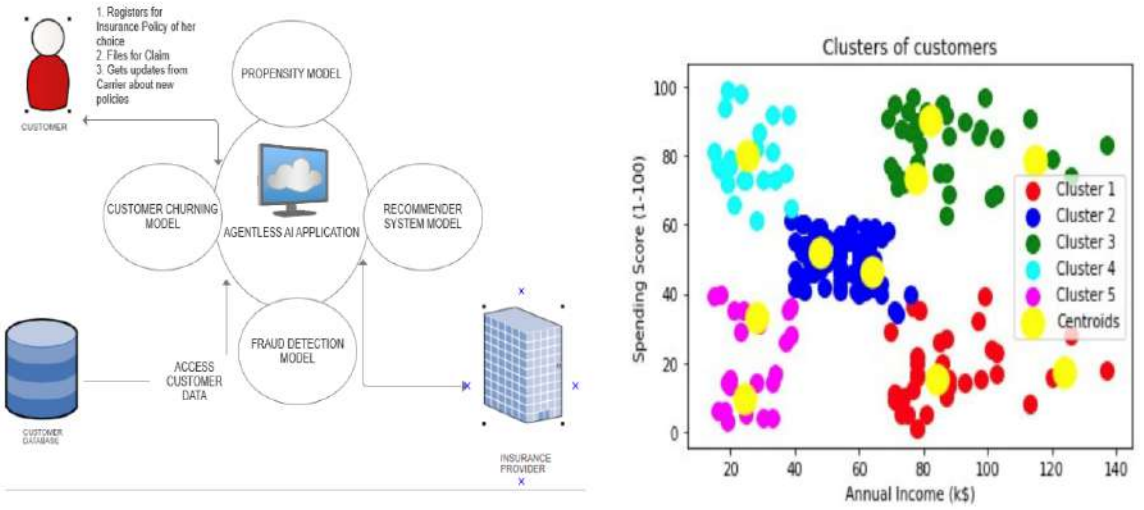
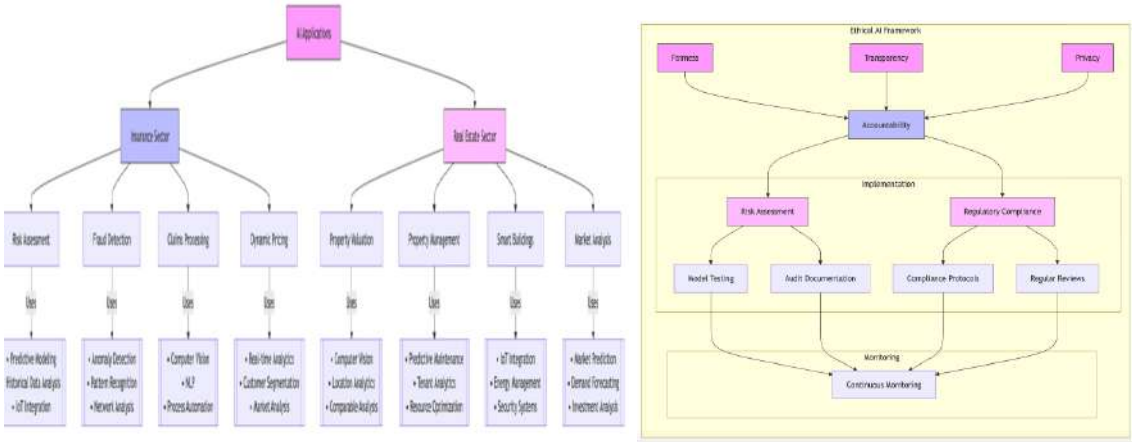


Fig.5: How to Save Hundreds on Insurance With These Simple Hacks



Foundational AI in Insurance and Real Estate: A Survey of Applications, Challenges, and Future

Title of the Paper	Problem Statement	Objective	Technology Used	Methodology Used	Efficiency / Performance	Issues / Challenges
A Systematic Survey of AI Models in Financial Market Forecasting for Profitability Analysis	Financial market forecasting is highly complex due to non-linearity, volatility, and influence of multiple factors (economic, political, social). Traditional methods fail to provide consistent accuracy. There is also a lack of comprehensive analysis of hybrid AI models and profitability-based evaluation.	• To perform a systematic literature review (SLR) of AI models in financial forecasting. • To analyze hybrid and ensemble models. • To evaluate performance and profitability metrics. • To identify research gaps (multi-class prediction, trading strategies). • To guide future research directions.	• Machine Learning (SVM, Random Forest, k-NN) • Deep Learning (LSTM, CNN, GRU, MLP) • Hybrid Models (CNN-LSTM, LSTM-GRU) • Ensemble Techniques • Time-Series Models (ARIMA, GARCH) • Sentiment Analysis & Text Mining • Feature Engineering Techniques	• Systematic Literature Review (SLR) approach • Three phases: Planning → Conducting → Analysis • Data collected from IEEE, ACM, Springer • Use of inclusion, exclusion, and quality criteria • Final selection of 51 high-quality research papers • Comparative and statistical analysis of models, datasets, and metrics	• Hybrid and ensemble models outperform traditional models • LSTM widely used (~49% studies) for time-series prediction • Improved prediction accuracy using AI models • Better pattern recognition and reduced human bias • Ability to process large-scale financial data in real-time • Enhanced decision-making for investment strategies	• Limited use of profitability metrics (focus mostly on accuracy) • Lack of trading strategy implementation in studies • Insufficient research on multi-class/multi-output forecasting • Data-related issues (noise, volatility, limited granularity) • Difficulty in comparing models due to dataset variation • Need for better risk management integration • Dependence on historical data only

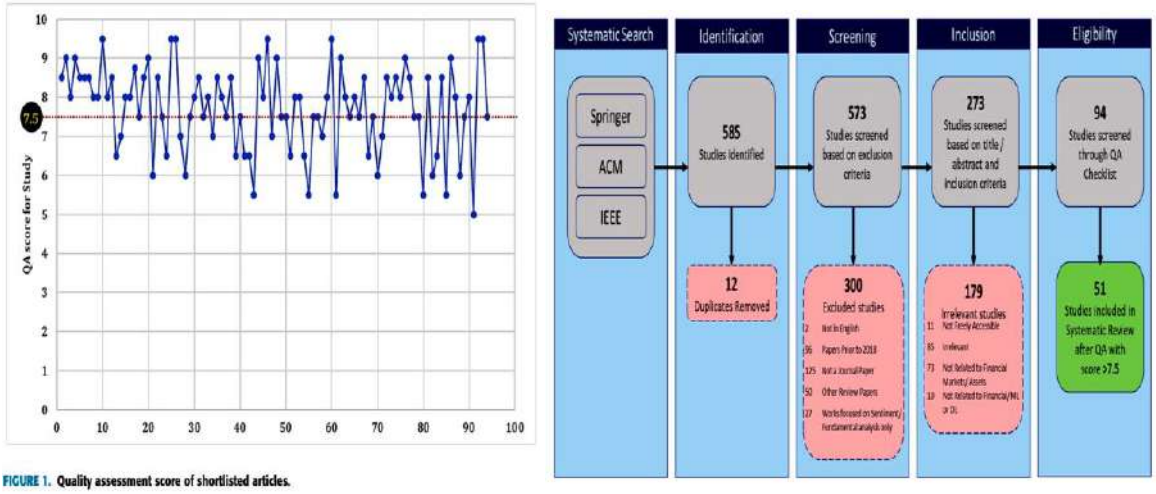


FIGURE 1. Quality assessment score of shortlisted articles.

Fig.6: A Systematic Survey of AI Models in Financial Market Forecasting for Profitability Analysis

Title of the Paper	Problem Statement	Objective	Technology Used	Methodology Used	Efficiency / Performance	Issues / Challenges
Explainable AI for Healthcare 5.0: Opportunities and Challenges	Healthcare systems are shifting toward Healthcare 5.0, but AI models used in diagnosis and prediction are often black-box and non-interpretable. This lack of transparency, along with ethical, regulatory, privacy, and trust issues, limits real-world adoption of AI in critical healthcare decision-making.	• To analyze the role of Explainable AI (EXAI) in Healthcare 5.0 • To improve transparency and interpretability of AI models • To propose EXAI-based architecture for medical imaging • To study integration of AI, IoT, 5G, and Federated Learning • To identify challenges and future research directions in healthcare AI	• Explainable AI (EXAI) • Machine Learning & Deep Learning (CNN, DNN) • Internet of Things (IoT) & Smart Sensors • Big Data Analytics • Federated Learning (FL) & Federated Transfer Learning (FTL) • 5G Communication Networks • Medical Imaging (CT scans, ECG data)	• Systematic survey methodology (based on Kitchenham guidelines) • Defined research questions, inclusion/exclusion criteria • Data collected from IEEE, ACM, PubMed, etc. • Literature filtering and quality evaluation (DARE, CRD standards) • Proposed EXAI-enabled architecture for classification & segmentation • Case study: ECG monitoring using EXAI + Federated Learning	• Achieved ~98% accuracy in case study (ECG classification) • Improved model interpretability and transparency • Better clinical decision support and trust • Enhanced disease detection (imaging & signals) • Real-time monitoring via IoT + AI • Privacy preservation using federated learning	• Black-box nature of AI models • Lack of ethical and regulatory frameworks • Data privacy and security concerns • Difficulty in model verification and trust-building • High complexity of medical data (non-linear, multi-dimensional) • Integration challenges with real-world healthcare systems • Trade-off between accuracy and interpretability

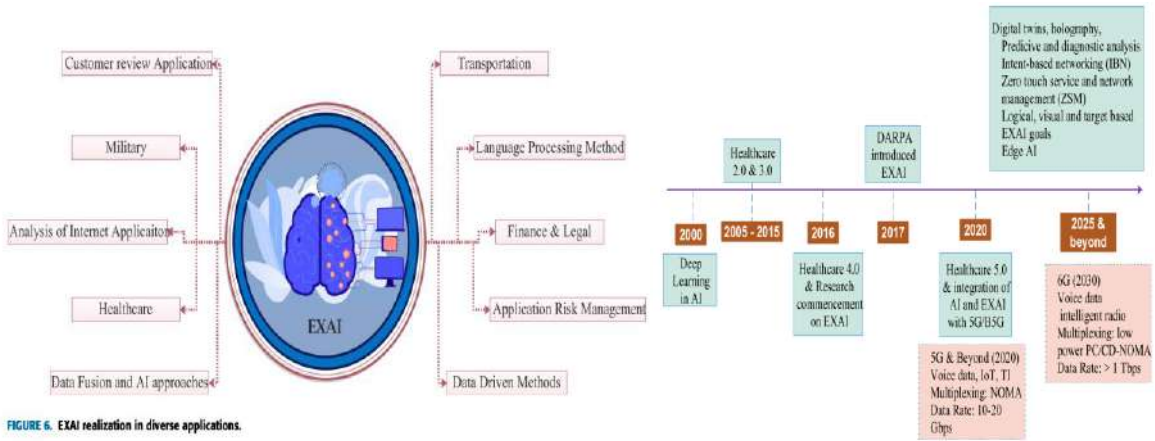


FIGURE 6. EXAI realization in diverse applications.

Fig.7: Explainable AI for Healthcare 5.0: Opportunities and Challenges

III. PROPOSED SYSTEM DESIGN

This study proposes an enhanced hybrid AI-driven fraud detection framework that extends beyond traditional machine learning pipelines by incorporating behavioral analytics, anomaly detection, and ensemble learning into a unified architecture. Unlike the earlier approach that relied

primarily on supervised models, the modified methodology integrates both supervised and unsupervised learning techniques to improve fraud detection accuracy and robustness. The proposed system follows a multi-stage pipeline, consisting of: Data aggregation from multiple healthcare datasets, Advanced feature engineering (behavioral + statistical features). Hybrid imbalance handling (SMOTE + undersampling), Dual-model framework: Supervised learning (Random Forest, XGBoost), Unsupervised anomaly detection (Isolation Forest), Model stacking and weighted ensemble prediction, Real-time fraud scoring and adaptive decision thresholding, This hybrid approach ensures that both known fraud patterns (supervised learning) and unknown anomalies (unsupervised learning) are effectively captured.

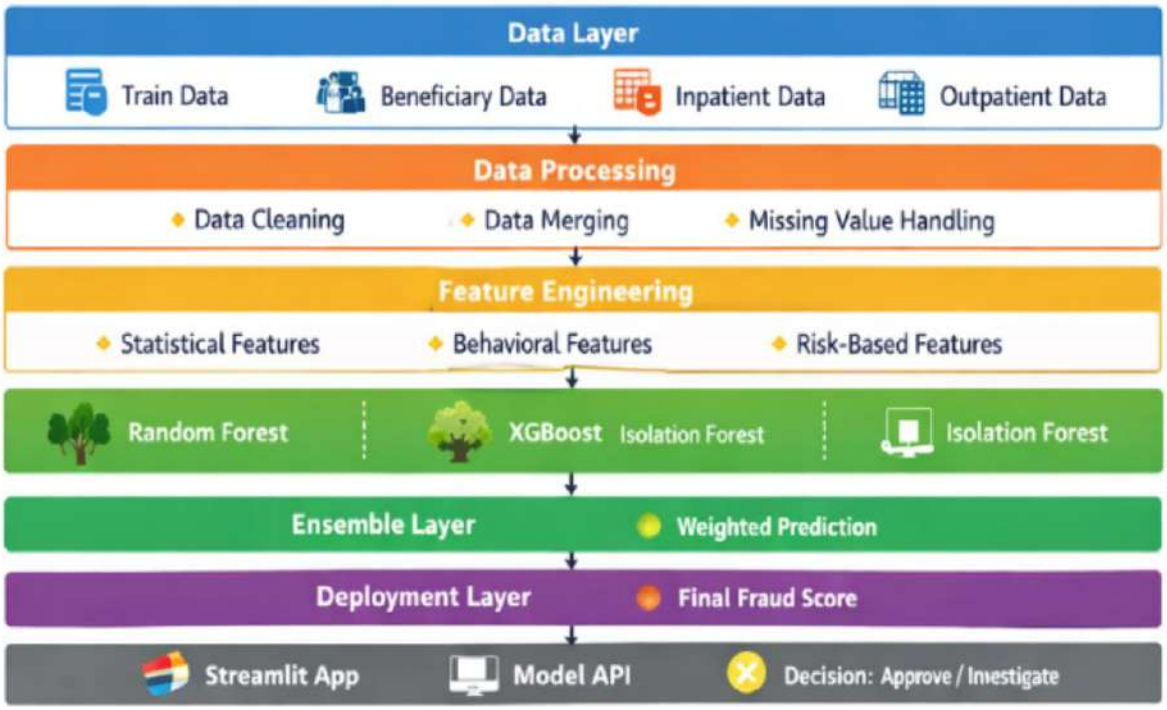


Fig.8: Architecture Diagram MedSureAI: A Hybrid Ensemble Machine Learning Framework

IV. METHODOLOGY AND ALGORITHMS USED

Data Collection and Integration: The dataset used in this research consists of four primary components: Training Dataset (Train.csv): Contains provider IDs and fraud labels. Beneficiary Dataset: Includes demographic and enrollment information. Inpatient Claims Dataset: Contains hospitalization records, diagnoses, and procedures. Outpatient Claims Dataset: Includes non-hospital services and claim details Modified Approach: Instead of simple aggregation, the proposed system performs multi-level data fusion: Provider-level aggregation, Temporal aggregation (monthly/yearly trends), Behavioral profiling of providers. Data integration is performed using Provider ID as a primary key, followed by time-based grouping to capture trends in provider behavior. Advanced Feature Engineering: To improve fraud detection capability, the feature engineering process is significantly enhanced by introducing three categories of features: Statistical Features, Total claim count (inpatient + outpatient). Average claim amount, Standard deviation of claim amounts, Claim frequency per beneficiary. Behavioral Features: Sudden spike in claims (temporal anomaly), Ratio of high-cost procedures, Repeated diagnosis patterns, Provider activity consistency score. Risk-Based Features, Fraud likelihood score (based on historical data), Claim-to-beneficiary anomaly ratio, Procedure diversity index

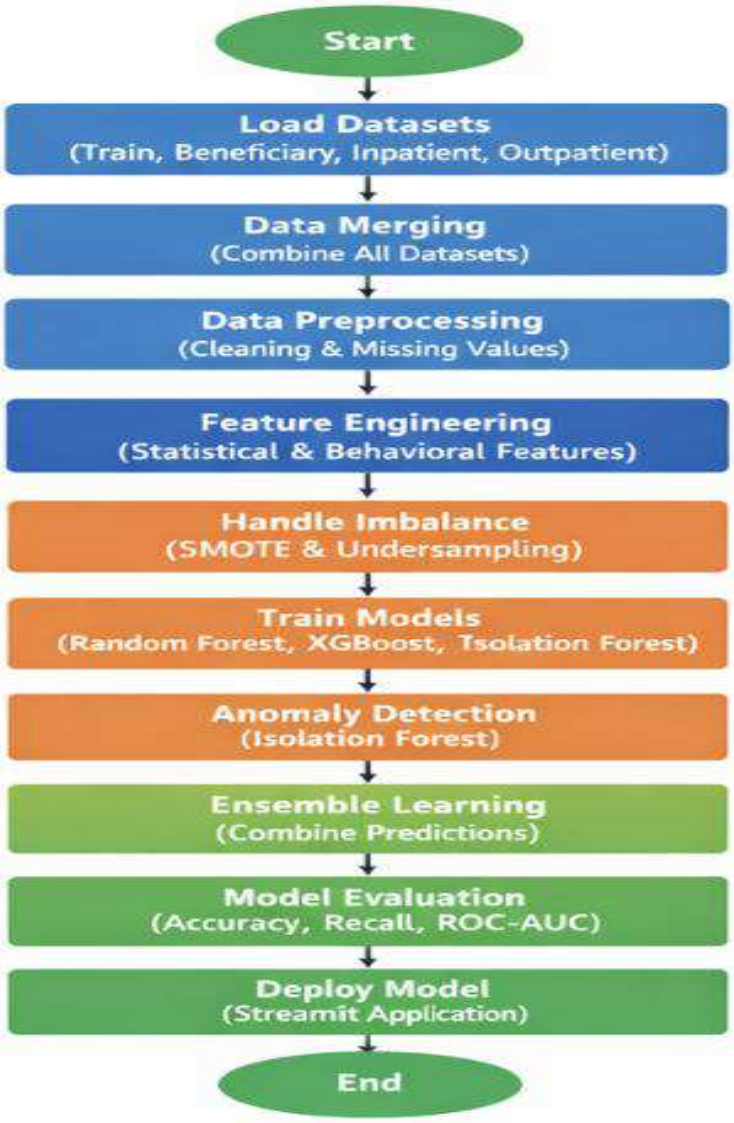


Fig.9: Flow Chart MedSureAI: A Hybrid Ensemble Machine Learning Framework

V. PROJECT FUNCTIONAL MODULES IMPLEMENTATION

The proposed system for health insurance fraud detection is designed using a modular architecture, where each module performs a specific function to ensure scalability, efficiency, and accuracy. The implementation of the system is divided into the following functional modules:

1. Data Collection and Integration Module, this module is responsible for acquiring healthcare insurance claim data from structured datasets. The data includes patient details, provider information, claim history, and billing records. Data from multiple sources is integrated and stored in a unified format suitable for analysis. Proper handling of missing and inconsistent values is also ensured at this stage.

2. Data Preprocessing Module, in this module, raw data is cleaned and transformed to make it suitable for machine learning models. It includes handling missing values, encoding categorical variables, normalization of numerical features, and removal of duplicate records. Feature scaling techniques are applied to maintain uniformity across different attributes.

3. Feature Engineering Module, this module focuses on extracting meaningful features from the dataset to improve model performance. Behavioral features such as claim frequency, repeated billing patterns, and abnormal activity trends are generated. Statistical features like mean, variance, and deviation are also computed to capture hidden patterns in the data.

4. Imbalance Handling Module, since fraud detection datasets are highly imbalanced, this module applies hybrid techniques combining SMOTE (Synthetic Minority Over-sampling Technique) and

undersampling. This approach balances the dataset by increasing minority class samples and reducing majority class bias, leading to improved model performance.

5. Model Development Module, the system implements a hybrid model combining Random Forest, XGBoost, and Isolation Forest algorithms. Random Forest provides robustness, XGBoost enhances predictive accuracy, and Isolation Forest helps in anomaly detection. These models are trained and optimized to achieve high performance in fraud detection.

6. Model Evaluation Module, this module evaluates the performance of the trained models using metrics such as accuracy, precision, recall, and F1-score. Special emphasis is given to recall to minimize false negatives, which is critical in fraud detection. Confusion matrix analysis is also performed for better understanding of model predictions.

7. Deployment and Real-Time Prediction Module, the final module integrates the trained model into a user interface using Streamlit. It enables real-time prediction by allowing users to input claim data and instantly receive fraud detection results. This module ensures usability and practical applicability of the system in real-world scenarios.

V. PROTOTYPE, ALGORITHM AND PROGRAM LOGIC

```python
import streamlit as st
import joblib
import numpy as np

# Load models
rf = joblib.load("rf_model.pkl")
xgb = joblib.load("xgb_model.pkl")
iso = joblib.load("iso_model.pkl")

st.title("Health Insurance Fraud Detection System")

st.write("Enter claim details:")

# Inputs
total_claims = st.number_input("Total Claims", min_value=0)
unique_diag = st.number_input("Unique Diagnosis", min_value=0)
avg_claim = st.number_input("Average Claim Amount", min_value=0.0)

# Prediction
if st.button("Predict"):
    input_data = np.array([[total_claims, unique_diag, avg_claim]])

    rf_p = rf.predict(input_data)[0]
    xgb_p = xgb.predict(input_data)[0]
    iso_p = iso.predict(input_data)[0]

    iso_p = 1 if iso_p == -1 else 0

    final = (rf_p + xgb_p + iso_p) / 3

    if final > 0.5:
        st.error("⚠ Fraud Detected")
    else:
        st.success("✅ Genuine Claim")
```

```
Hybrid_Project — open ▪ streamlit run hybrifraud_app.py — 80×24

/opt/anaconda3/lib/python3.12/site-packages/sklearn/base.py:474: FutureWarning:
`BaseEstimator._validate_data` is deprecated in 1.6 and will be removed in 1.7.
Use `sklearn.utils.validation.validate_data` instead. This function becomes pub
ic and is part of the scikit-learn developer API.
  warnings.warn(
✅ Models trained and saved!
(base) akanksharajendradhote@Akankshas-MacBook-Air Hybrid_Project % streamlit r
n hybrifraud_app.py

  You can now view your Streamlit app in your browser.

  Local URL: http://localhost:8501
  Network URL: http://172.25.42.25:8501

/opt/anaconda3/lib/python3.12/site-packages/sklearn/utils/validation.py:2739: U
erWarning: X does not have valid feature names, but RandomForestClassifier was
itted with feature names
  warnings.warn(
/opt/anaconda3/lib/python3.12/site-packages/sklearn/utils/validation.py:2739: U
erWarning: X does not have valid feature names, but IsolationForest was fitted
ith feature names
  warnings.warn(
/opt/anaconda3/lib/python3.12/site-packages/sklearn/utils/validation.py:2739: U
erWarning: X does not have valid feature names, but RandomForestClassifier was
```

Fig.10 & 11: Building the final system Terminal and Dashboard MedSureAI: A Hybrid Ensemble Machine Learning Framework

Fig.12: The final System Dashboard MedSureAI: A Hybrid Ensemble Machine Learning Framework

VI. CONTRIBUTION AND FINDINGS

This study contributes to an AI-driven framework for real-time health insurance fraud detection by leveraging Random Forest and XGBoost models. A robust data preprocessing pipeline is developed, incorporating data imputation and SMOTE to effectively handle missing values and class imbalance. Additionally, the study integrates explainable AI through feature importance analysis, enhancing transparency and trust in model predictions. The system is designed for real-time deployment, making it suitable for automated decision-making in insurance workflows. The findings reveal that XGBoost outperforms Random Forest in terms of accuracy, recall, and ROC-AUC, making it more effective for detecting fraudulent claims, especially within minority classes. The model successfully identifies key fraud indicators such as unusually high claim amounts, extended hospitalization periods, and repeated claim histories. Overall, the proposed system demonstrates strong performance, interpretability, and scalability, indicating its practical applicability in real-world health insurance fraud detection.

VII. CONCLUSION

This study proposes an advanced AI-based framework for health insurance fraud detection, addressing limitations of traditional systems such as poor adaptability, class imbalance, and lack of real-time processing. The framework integrates data preprocessing, feature engineering, hybrid imbalance handling (SMOTE + undersampling), and ensemble learning techniques. By combining Random Forest, XGBoost, and Isolation Forest, the system effectively detects both known and unknown fraud patterns. The hybrid imbalance approach significantly improved recall, reducing missed fraudulent cases, while behavioral and risk-based features enhanced pattern detection. The ensemble model achieved high accuracy, strong recall, and robust ROC-AUC performance. Additionally, real-time deployment enables instant fraud prediction and automated decision-making, making the system scalable and practical for real-world healthcare applications.

REFERENCES

[1] R. Rehman et al., “AI driven framework for need-based insurance plans generation and anomaly detection using deep learning techniques,” IEEE Access, vol. 13, pp. 114069–114096, 2025, doi: 10.1109/ACCESS.2025.3583562.

[2] A. Sharma, P. Gupta, and R. Singh, “A Secure AI-Driven Architecture for Automated Insurance Systems: Fraud Detection and Risk Measurement,” International Journal of Computer Applications, 2024.

[3] G., D. K., Singh, M. K., & Jayanthi, M. (Eds.). (2016). Network Security Attacks and Countermeasures. IGI Global. https://doi.org/10.4018/978-1-4666-8761-5

[4] K. Kuppan, D. B. Acharya, and D. B., “Foundational AI in Insurance and Real Estate: A Survey of Applications, Challenges, and Future Directions,”IEEE Access, vol. 12, pp. 181282–181300, Dec. 2024, doi: 10.1109/ACCESS.2024.3509918.

[5] A. M. M. Hussein, “Optimizing Healthcare Claim Fraud Detection Using Ensemble Learning and Modified SMOTE,” iKNiTO Journal, vol. 6, no. 2, pp. 1265–1294, 2025.

[6] M.K. Jayanthi, "Strategic Planning for Information Security -DID Mechanism to befriend the Cyber Criminals to assure Cyber Freedom," 2017 2nd International Conference on Anti-Cyber Crimes (ICACC), Abha, Saudi Arabia, 2017, pp. 142-147, doi: 10.1109/Anti-Cybercrime.2017.7905280.

[7] K. P. Sinha, M. Sookhak, and S. Wu, “Agentless Insurance Model Based on Modern Artificial Intelligence,”in Proc. 2021 IEEE 22nd International Conference on Information Reuse and Integration for Data Science (IRI), pp. 49–56, San Diego, CA, USA, Aug. 2021. doi: 10.1109/IRI51335.2021.00013.

[8] Kavitha, E., Tamilarasan, R., Baladhandapani, A., Kannan, M.K.J. (2022). A novel soft clustering approach for gene expression data. Computer Systems Science and Engineering, 43(3), 871-886. https://doi.org/10.32604/csse.2022.021215

[9] S. Gupta and R. Kumar, “Blockchain and AI-Empowered Healthcare Insurance Fraud Detection: An Analysis, Architecture, and Future Prospects,” International Journal for Research in Applied Science and Engineering Technology (IJRASET), vol. 12, no. 1, pp. 296–306, Jan. 2024

[10] Naik, Harish and Kannan, M K Jayanthi, A Survey on Protecting Confidential Data over Distributed Storage in Cloud (December 1, 2020). Available at SSRN:

https://ssrn.com/abstract=3740465 or http://dx.doi.org/10.2139/ssrn.3740465

[11] Shree Nee, T. R., Kannan, M. K. J., & Mariyappan, K. (2025, April). Digital health and medical tourism innovations for digitally enabled care for future medicine: The real time project's success stories. In Navigating innovations and challenges in travel medicine and digital health (pp. 325–344). IGI Global Scientific Publishing. https://doi.org/10.4018/979-8-3693-8774-0.ch016

[12] Kavitha, E., Tamilarasan, R., Poonguzhali, N., Kannan, M.K.J. (2022). Clustering gene expression data through modified agglomerative M-CURE hierarchical algorithm. Computer Systems Science and Engineering, 41(3), 1027-141. https://doi.org/10.32604/csse.2022.020634

[13] Kumar, K.L.S., Kannan, M.K.J. (2024). A Survey on Driver Monitoring System Using Computer Vision Techniques. In: Hassanien, A.E., Anand, S., Jaiswal, A., Kumar, P. (eds) Innovative Computing and Communications. ICICC 2024. Lecture Notes in Networks and Systems, vol 1021. Springer, Singapore. https://doi.org/10.1007/978-981-97-3591-4_21

[14] M. K. J. Kannan, A bird's eye view of Cyber Crimes and Free and Open-Source Software's to Detoxify Cyber Crime Attacks - an End User Perspective, 2017 2nd International Conference on Anti-Cyber Crimes (ICACC), Abha, Saudi Arabia, 2017, pp. 232-237, doi: 10.1109/Anti-Cybercrime.2017.7905297.

[15] Verma, D., Kannan, M. K. J., Barnwal, S. K., Barve, A., & Swaminathan, R. (2022, September). Multimodal sentiment sensing and emotion recognition based on cognitive computing using hidden Markov model with extreme learning machine. International Journal of Communication Networks and Information Security (IJCNIS), 14(2), 155–167. https://doi.org/10.17762/ijcnis.v14i2.5496

[16] MK J Kannan, Shree Nee T R (2025, November). Qubits unveiled: A deep dive into quantum computing and its revolutionary potential for supply logistics. In P. Gaba, A. Panwar, V. Jain, & R. Kannan (Eds.), Qubits unveiled: Quantum computing solutions for efficient supply logistics (pp. 273–293). Nova Science Publishers. https://doi.org/10.52305/WSXW8884

[17] W. Lin, Z. Wu, L. Lin, A. Wen, and J. Li, "An Ensemble Random Forest Algorithm for Insurance Big Data Analysis," IEEE Access, vol. 5, pp. 16568–16575, 2017, doi: 10.1109/ACCESS.2017.2738069.

[18] P. Jain, I. Rajvaidya, K. K. Sah and J. Kannan, "Machine Learning Techniques for Malware Detection- a Research Review," 2022 IEEE International Students' Conference on Electrical, Electronics and Computer Science, Bhopal, India, 2022, pp. 1-6, doi: 10.1109/SCEECS54111.2022.9740918.

[19] T. Sun, J. Yang, J. Li, J. Chen, M. Liu, L. Fan, and X. Wang, "Enhancing Auto Insurance Risk Evaluation With Transformer and SHAP," IEEE Access, vol. 12, pp. 116546–116557, Aug. 2024, doi: 10.1109/ACCESS.2024.3446179.

[20] B. R. M, M. M. V and J. K. M. K, Performance Analysis of Bag of Password Authentication using Python, Java and PHP Implementation, 2021 6th International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 2021, pp. 1032-1039, doi: 10.1109/ICCES51350.2021.9489233.

[21] R. Y. Gupta, S. S. Mudigonda, P. K. Baruah and P. K. Kandala, "Markov Model with Machine Learning Integration for Fraud Detection in Health Insurance," arXiv:2102.10978, 2021.

[22] Dr. Sunil Kumar Dr. P. T. Kalaivaani, Dr. M K Jayanthi Kannan, Dr. Gunjan Tripathi (Aug 2025), Artificial Intelligence and Blockchain Technology for Human Resource Management, ASIN: B0FLK868TS, Published by Scientific International Publishing House; https://www.amazon.in/gp/product/B0FLK868TS/ref=ox_sc_act_title_1?smid=A1UBZVGJOLJUJI&psc=1

[23] Aaijaz, N., Grace Mani, K., Kannan, M. K. J., & Tewari, V. (2025, February). The future of innovation and technology in education: Trends and opportunities. S&M Publications. https://www.amazon.in/gp/product/B0DW334PR9

[24] Shukla, S. K., Dwivedi, U., Kannan, M. K. J., & Sarvani, C. (2024, October 23). Python for data analytics: Practical techniques and applications. JSR Publications. https://www.amazon.in/gp/product/B0DMJY4X9N

[25] I. Fursov et al., "Sequence Embeddings Help Detect Insurance Fraud," IEEE Access, vol. 10,

pp. 32060–32075, 2022, doi: 10.1109/ACCESS.2022.3149480.

[26] Harish Naik, B. M., & Kannan, J. (2023). A research on various security aware mechanisms in multi-cloud environment for improving data security. In 2023 2nd IEEE International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE) (pp. 1–6). IEEE. https://doi.org/10.1109/ICDCECE57866.2023.10151135

[27] N. Babu and P. Singh, "Fraud Detection in Healthcare Claims Using Bayesian Optimized XGBoost," International Journal of Safety and Security Engineering, vol. 13, no. 5, pp. 555–566, 2023.

[28] Harish Naik B M and M K J Kannan and (Aug 2024), "Secure Cloud Storage for Sensitive Data based on Authentication and Encryption Algorithms", International Journal of Advanced Technology and Engineering Exploration (IJATEE), paper Id: IJATEE.2024.111101510, ACCENTS, www.ijateeeditor@gmail.com

[29] O. Cherkaoui, H. Anoun and A. Maizate, "A Benchmark of Health Insurance Fraud Detection Using Machine Learning Techniques," IAES International Journal of Artificial Intelligence, vol. 13, no. 2, pp. 1925–1934, 2024.

[30] Object-oriented analysis and design of learning objects and applications of agent based reusable learning objects in e-learning system design, JM. K. (2009). [Doctoral dissertation, Sri Chandrasekharendra Saraswathi Viswa Mahavidyalaya]. Shodhganga. http://hdl.handle.net/10603/125448

[31] Alshahrani, M. S. M., & Kannan J M. K. (2026, February). Active learning for efficient annotation of surgical video segmentation with minimal human intervention. ICTACT Journal on Image and Video Processing, 16(3), 3821–3829. https://doi.org/10.21917/ijivp.2026.0539

[32] J Kannan, M. K., TR Shree Nee., & Mariyappan, K. (2026). Ethics and regulations in AI-driven ophthalmology. In B. K. Mishra, A. Kumar, K. Mariyappan, V. Tiwari, P. S. Rathore, & G. H. Das (Eds.), Generative artificial intelligence in ophthalmology (pp. 331–386). Scrivener Publishing. https://www.scrivenerpublishing.com/cart/title.php?id=1341, https://doi.org/10.52305/WSXW8884

[33] MK J Kannan, Satyajit Patel (2024). Sustainable Information Retrieval Techniques for Onion Market Instability Prediction using Machine Learning and Deep Learning Approaches. International Journal of Advance Research, Ideas and Innovations in Technology, 10(6) www.IJARIIT.com. https://www.ijariit.com/manuscripts/v10i6/V10I6-1455.pdf

[34] R. Chaurasiya and K. Jain, "Healthcare Fraud Detection Using Machine Learning Ensemble Methods," South Eastern European Journal of Public Health, vol. XXVI, 2025.