Notification No.: 576/2025

Date of Award: 17-03-2025

Name of Scholar: Lavina Pahuja

Name of Supervisor: Dr. Ahmad Kamal

Name of Department: Department of Mathematics

Topic of Research: "Artificial Intelligent Approaches for Web-based Fraud Detection"

Findings

This thesis addresses the challenges of web-based fraud detection by proposing AI-driven machine learning approaches that leverage advanced feature engineering and classification algorithms to effectively identify fraudulent activities. The proposed methods achieve high accuracy across various fraud types, including Ethereum cryptocurrency, online dating, and healthcare insurance fraud, providing valuable tools for safeguarding individuals and organizations. The CRISP-DM framework is utilized for detecting fraudulent transactions on the Ethereum blockchain, where a comparative analysis of ten machine learning models demonstrates superior performance of ensemble classifiers, with LightGBM achieving 99.2% accuracy.

For detecting fake profiles, the thesis integrates a modified LightGBM algorithm with SMOTE within the OSEMN framework, significantly improving classification accuracy by incorporating syntactic, semantic, linguistic, and user-based features. A scalable approach using enhanced K-means clustering is also proposed for profiling suspicious users on online platforms, combining unsupervised clustering with supervised validation.

In healthcare insurance fraud detection, the thesis introduces NetSLFD, a network-based stacked learning model that captures complex relationships among healthcare entities. Evaluated on a public dataset, NetSLFD achieves 99.38% accuracy, outperforming existing methods.

Overall, this research contributes to the development of robust fraud detection systems by integrating machine learning, text analytics, and network-based methodologies, significantly enhancing fraud prevention across online domains. The thesis is structured into seven chapters, summarized as follows:

Chapter 1 provides an overview of frauds and their types, focusing on web-based fraud, and highlights the role of AI in fraud detection. It outlines the challenges, objectives, and structure of the thesis.

Chapter 2 examines the evolution of fraud detection techniques, with an emphasis on AI-driven methods for detecting Ethereum cryptocurrency, online dating, and healthcare insurance frauds. It reviews existing studies, showing the transition from traditional statistical methods to advanced AI solutions and identifies research gaps in the field.

Chapter 3 introduces EnLEFD-DM, a classification model for Ethereum blockchain fraud detection within the CRISP-DM framework, for detecting fraudulent Ethereum blockchain transactions. The model uses an optimized feature set and SMOTE to address dataset imbalance, with LightGBM achieving 99.2% accuracy. Experimental results demonstrate the superiority of this approach over existing methods.

Chapter 4 discusses the integration of unstructured text data from user profiles with structured data to detect fake accounts. It presents a novel approach that enhances the LightGBM algorithm and incorporates SMOTE within the Obtain, Scrub, Explore, Model, and Interpret (OSEMN) framework. Advanced feature engineering techniques were utilized to develop hybrid feature sets, incorporating syntactic, semantic, linguistic, and user-related attributes. Multiple ensemble learning models were assessed, with the enhanced LightGBM algorithm delivering the best performance, achieving 98.6% accuracy. A comparative evaluation using the MIB Twitter dataset highlights its effectiveness across different online platforms.

Chapter 5 introduces a hybrid method that combines multiple machine learning techniques, with an emphasis on enhanced k-means clustering for classifying unlabeled data. The approach begins with advanced feature selection and cluster optimization to form user profile clusters, which are then evaluated using internal validation metrics. In the next phase, supervised learning techniques validate these clusters, followed by testing on a new dataset. This hybrid framework demonstrates that combining clustering with supervised validation significantly improves the accuracy and reliability of profile detection. By leveraging a comprehensive dataset and employing the Minkowski distance metric, the analysis identifies distinct patterns of suspicious behavior, providing valuable domain-specific insights.

Chapter 6 proposes NetSLFD, an innovative approach for healthcare insurance fraud detection that combines network analysis with stacked machine learning. It models interactions among providers, beneficiaries, and physicians using a tripartite graph and extracts network-based features, including centrality measures and community detection metrics. These features are combined with domain-specific and derived attributes and processed through a stacked ensemble model comprising XGBoost, LightGBM, and Random Forest as base classifiers, with a Multilayer Perceptron as the meta-learner. The approach is evaluated on a publicly available healthcare dataset, achieving an accuracy of 92.52% and demonstrating its applicability for fraud detection beyond healthcare.

Chapter 7 concludes the thesis by summarizing its contributions, addressing challenges encountered, and suggesting future research directions in the evolving field of fraud detection. It also highlights important considerations for future advancements.