STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **14**, December 2025, pp 3565–3588.
Published online in International Academic Press (www.IAPress.org)

Enhancing Fraud Detection in Health Insurance: Deep Neural Network Approaches and Performance Analysis

Gaber Sallam Salem Abdalla [1,*], Mohamed F. Abouelenein [1], Hatem M. Noaman [2]

[1] *Department of Insurance and Risk Management, College of Business, Imam Mohammad Ibn Saud Islamic University, Saudi Arabia*
[2] *Department of Computer Science, Faculty of Computer Science and Artificial Intelligence, Beni-Suef University, Beni-Suef 62511, Egypt*

Abstract This study develops and examines a comprehensive deep learning framework for the detection of multi-class healthcare fraud in National Health Insurance Scheme (NHIS) claims. We examined 20,388 NHIS healthcare claims revealing four specific fraud patterns: Phantom Billing, Wrong Diagnosis, Ghost Enrollee, and legitimate claims. Four different deep neural network architectures were developed and evaluated: Simple NN, Deep Wide NN, Regularized NN, and Residual NN, in addition to ensemble methods. The Simple Neural Network achieved the highest overall performance, with a test accuracy of 79.84% and an F1-macro score of 77.76%. Despite possessing only 100,324 parameters (five times fewer than the Wide Deep Neural Network), it outperformed more complex designs while achieving the fastest training time of 40.61 seconds. Multiclass analysis demonstrated exceptional performance in Ghost Enrollee detection (97.84% F1-score) and moderate performance in Phantom Billing detection (61.15% F1-score).

Keywords Deep Learning, Multi-Class Classification, Healthcare Fraud Detection, Neural Networks, Feature Engineering, NHIS Claims

DOI: 10.19139/soic-2310-5070-3097

1. Introduction

1.1. Background

Healthcare fraud represents an increasing and serious risk to global healthcare systems, resulting in considerable revenue losses and lowering the integrity of healthcare services and the quality of care provided to patients. Estimates suggest that 3%-10% of total spending on healthcare is wasted due to fraudulent activities [1]. In the United States, healthcare insurance fraud, especially Medicare and Medicaid fraud, is the most expensive type of insurance fraud. Settlements and judgments related to healthcare fraud under the False Claims Act exceed $1.32 trillion annually, accounting for 30% of total spending on healthcare [2]. The evolution of healthcare fraud schemes has become more complex, incorporating various fraudulent activities that exploit vulnerabilities within the healthcare system [3]. Conventional binary classification methods (fraud vs. non-fraud) do not sufficiently address the complexities of various fraud types, thereby constraining the effectiveness of the prevention strategies. Multiclass fraud detection allows healthcare organizations to implement specialized measures for specific types of fraud, optimize resource allocation, and enhance prevention strategies.

1.2. Problem Statement

Healthcare fraud detection presents distinct challenges that set it apart from other fraud detection domains. The class imbalance problem, defined by a lack of fraudulent cases compared to legitimate claims, results in an

*Correspondence to: Gaber Sallam Salem Abdalla (Email: : jssabdullah@imamu.edu.sa). Department of Insurance and Risk Management, College of Business, Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh 11432, Riyadh, Saudi Arabia.

ISSN 2310-5070 (online) ISSN 2311-004X (print)

algorithmic bias that chooses the majority class [4]. The complexity of medical coding systems, high-dimensional mixed data types, and privacy constraints further complicate detection challenges [5]. Recent studies emphasize the necessity of separating various types of fraud instead of categorizing fraud as a single entity. Phantom billing refers to charges for services that are not provided. Ghost Enrollee fraud pertains to the use of fake patient identities, while Wrong Diagnosis fraud involves planned misclassification to secure higher payments. Different types of fraud require distinct detection strategies and prevention measures. The utilization of deep learning in healthcare fraud detection has progressed from fundamental rule-based systems to advanced neural networks that can identify complex patterns within complex claims datasets [1]. Most existing research primarily addresses binary classification, which restricts the practical applicability of healthcare administrators requiring detailed insights into fraud types for effective intervention.

1.3. Research Gap and Objectives

This study makes significant contributions to the field of healthcare fraud detection. This study presents a comprehensive multi-class deep learning evaluation customized for healthcare fraud categorization, advancing from binary fraud/no-fraud classification to facilitate targeted intervention strategies. The confidence-based deployment framework provides a viable approach that combines automation efficiency with the necessity for human oversight, thereby addressing a significant gap in the implementation of real-world fraud detection systems. A comprehensive analysis of overfitting and generalization offers important insights into model selection criteria for fraud detection applications, indicating that simpler architectures may provide greater reliability. This paper is structured as follows: Literature Review in Section 2 presents a detailed discussion of fraud detection in healthcare claims and suggested approaches and related work for this problem. The Materials and Methods section details the dataset characteristics and deep learning approaches employed in this evaluation. The Results and Discussion section presents the evaluation outcomes, including performance metrics, confusion matrix analysis, misclassification patterns, model calibration analysis, and feature importance findings. The Conclusion section synthesizes the essential findings, addresses the limitations, and provides recommendations for future research and the practical implementation of machine learning-based healthcare fraud detection systems.

2. Related Work

The integration of machine learning and artificial intelligence techniques in fraud detection in healthcare and insurance systems has received considerable academic focus in the last ten years. This research encompasses multiple domains, such as credit card fraud detection and healthcare insurance fraud prevention, utilizing various methodologies and tackling the technical challenges associated with fraud detection systems.

2.1. Machine Learning Approaches

Research on fraud detection primarily emphasizes traditional machine learning algorithms within the financial and healthcare sectors. Varmedja et al. (2019) established foundational benchmarks through the evaluation of Logistic Regression, Random Forest, Naive Bayes, and Multilayer Perceptron models in the context of credit card fraud detection. The findings demonstrated that Random Forest exhibited enhanced performance, achieving a precision of 96.38% and a recall of 81.63%, especially when supplemented with SMOTE oversampling techniques [1]. This study highlights the significance of addressing class imbalance, a recurring issue in the fraud detection literature. Several studies have explored traditional machine learning applications within healthcare contexts, building on these foundational principles. Nabrawi and Alanazi (2023) illustrated the effectiveness of Random Forest in analyzing healthcare insurance claims, emphasizing demographic insights and the interpretability of ensemble methods [5]. Severino and Peng (2021) performed comparative analyses of various algorithms, such as Logistic Regression, Random Forest, Gradient Boosting Machines, XGBoost, and LightGBM, in the context of property insurance fraud. The study concluded that ensemble methods consistently surpassed single classifiers when suitable preprocessing techniques were utilized [6]. Prova (2024) developed a comprehensive fraud detection system addressing healthcare fraud in the United States, utilizing a combination of traditional machine learning

algorithms, ensemble methods, and deep learning techniques. The study employed a dataset of 558,211 records with 55 demographic, medical, and financial variables. The Stacking Ensemble achieved the best performance with an accuracy of 92.79% and an ROC AUC of 96.95%, while XGBoost achieved the highest precision at 97.34%. The research incorporated SHAP value analysis for interpretability and developed a real-time claims processing system with automated retraining mechanisms [7]. Bounab et al. (2024) addressed the critical challenge of class imbalance in Medicare Part B fraud detection by proposing a hybrid resampling approach combining SMOTE with Edited Nearest Neighbors (SMOTE-ENN). Using a 2020 Medicare Part B dataset of over 9.4 million records, the method was tested with six machine learning classifiers. Results showed that SMOTE-ENN significantly improved minority class detection, with Decision Trees achieving near-perfect performance metrics and consistently high AUC and AUPRC values [8]. Sumalatha and Prabha (2019) advanced this field by investigating the integration of Logistic Regression with multi-criteria decision analysis for managing mediclaim fraud [9].

2.2. Deep Learning Approaches

Recent studies demonstrate a clear evolution toward more sophisticated methodological approaches by integrating ensemble learning with deep learning techniques. Gupta et al. (2021) performed a detailed comparative analysis of machine learning and deep learning models in the context of universal health coverage schemes. Their findings indicated that neural networks trained on undersampled data attained F1-scores of 0.95, whereas Gradient Boosting Machines integrated with Tabular GANs exhibited strong performance across various evaluation metrics [10]. Wang et al. (2025) proposed an ensemble framework that integrates XGBoost, Random Forest, and Logistic Regression, augmented with SHAP (SHapley Additive exPlanations) to enhance interpretability [11]. This approach meets the essential requirement for explainable fraud detection systems, emphasizing the importance of transparency and trust in healthcare decision-making processes. Johnson and Khoshgoftaar (2019) offered complementary insights, showing that neural networks can compete with traditional methods when trained on balanced datasets. However, they highlighted the persistent challenges associated with real-world imbalanced fraud data [12]. Matloob et al. (2025) introduced an innovative framework combining machine learning and deep learning to detect fraud at the level of patients, providers, and services. Their two-stage approach utilizes an association rule engine for detecting suspicious transactions and an Anomaly Transformer for examining time-series service patterns. The framework demonstrated effectiveness in categorizing fraud as patient-level (50%), service vs. doctor (12%), service vs. patient (13%), and physician-level (25%) [13]. Shah et al. (2022) conducted a comprehensive evaluation of machine and deep learning techniques for financial fraud detection in the healthcare industry, focusing on credit card fraud. Their study compared Naive Bayes, Logistic Regression, KNN, Random Forest, CNN, and deep ANN algorithms. Results showed that deep ANN achieved the best performance with 98.53% accuracy, 96.74% precision, 94.52% recall, and 97.10% F1-score, demonstrating the superiority of deep learning in capturing complex fraud patterns [14]. Shungube et al. (2024) evaluated three deep learning models—Artificial Neural Networks (ANN), Convolutional Neural Networks (CNN), and Long Short-Term Memory (LSTM) networks—on a real Medicare healthcare claims dataset. The ANN achieved the best overall performance with 94% accuracy, 0.78 precision, and an F1-score of 0.57, while CNN excelled in reducing false positives and LSTM was better at capturing temporal patterns. The study enhanced interpretability using LIME analysis to identify important prediction factors [15]. Anand Kumar and Sountharrajan (2025) proposed a deep 12-layer CNN model optimized by the Enhanced Hippopotamus Optimization Algorithm (EHOA) for insurance claims estimation and fraud detection. The EHOA-CNN-12 achieved 92% accuracy and outperformed baseline models including VGG16, VGG19, and ResNet50 by 3-7% in various metrics. The approach demonstrated faster convergence, reduced overfitting, and strong potential for real-time fraud detection [16]. Suesserman et al. (2023) addressed procedure code overutilization detection using unsupervised deep learning methods. Their study employed autoencoders with a novel feature-weighted binary cross-entropy loss function to handle sparse, imbalanced data. The autoencoder significantly outperformed baseline DBSCAN clustering, achieving high recall and F1-scores (F1-score 0.97 on synthetic data; 0.63 on manually annotated claims) [17].

2.3. Graph-Based and Network Analysis Approaches

An innovative research trajectory involves the use of graph and network analytics for fraud detection. Zhou et al. (2023) introduced FraudAuditor, a visual analytics system designed to detect collusive fraud involving patients, medical institutions, and drugstores through co-visit network analysis and community detection algorithms [18]. This study underscores the significance of modeling complex relationships between entities in healthcare fraud scenarios. Yoo et al. (2023) conducted a comprehensive comparative analysis of traditional machine learning and Graph Neural Networks (GNNs) for Medicare fraud detection, transforming claims data into heterogeneous graph structures where nodes represent providers, patients, and medical services [19]. Their findings revealed that while classical ensemble models, such as CatBoost, achieved AUC-ROC scores of 0.91, GraphSAGE demonstrated competitive performance at 0.87, highlighting the potential of graph-based modeling for capturing relational information without extensive feature engineering. Branting et al. (2016) contributed to this domain by developing graph-analytic approaches for healthcare fraud risk estimation, achieving 82% accuracy in identifying known fraudulent providers through structural similarity analysis and behavioral pattern recognition [20]. These graph-based approaches represent a paradigm shift from traditional feature-based models to relationship-aware fraud-detection systems.

2.4. Domain-Specific Applications

Several studies have explored specialized techniques tailored to specific fraud-detection challenges. Johnson and Khoshgoftaar (2021) investigated medical provider embeddings using dense vector representations learned from historical claims data, though they noted limitations requiring larger datasets and more sophisticated embedding techniques [21]. Nugraha et al. (2022) proposed GAN-based oversampling techniques as superior alternatives to conventional methods like SMOTE for addressing severe class imbalance in healthcare fraud detection [22]. Sadiq and Shyu (2019) introduced a cascaded propensity matching framework that leverages propensity score matching and concept drift learning to handle evolving fraud patterns in Medicare programs, demonstrating improved sensitivity and recall in detecting new fraud practices [23]. This work addresses the dynamic nature of fraud schemes and the need for adaptive detection systems.

2.5. Data-Centric and Systematic Approaches

Recent research has investigated the importance of data-centric methodologies and systematic evaluation techniques. Johnson and Khoshgoftaar (2023) concentrated on advancing fraud detection through data-centric AI, introducing six novel labeled datasets enriched with provider-level, claims-level, and beneficiary-level statistics, which demonstrated significant performance enhancements when utilizing aggregated-enriched datasets [24]. Hancock and Khoshgoftaar (2021) explored Gradient Boosted Decision Trees, systematically evaluating classifier selection and sampling strategies, and confirmed the statistical significance of these factors on model performance [25]. Du Preez et al. (2024) conducted a comprehensive systematic review of machine learning techniques for healthcare fraud detection, identifying a lack of standardization in data preprocessing, feature engineering, and evaluation protocols across studies [4]. This review highlights the necessity for unified frameworks and standardized evaluation methodologies to facilitate fair comparisons and enhance reproducibility.

2.6. Feature Engineering and Explainability

The importance of explainable artificial intelligence (AI) in fraud detection has been consistently emphasized in numerous studies. Hancock et al. (2023) developed ensemble-based supervised feature selection methods that improved model interpretability while maintaining high performance, with XGBoost achieving AUPRC scores of 0.9408 for Medicare fraud detection [2]. Zhang et al. (2020) proposed hybrid frameworks that integrate rule-based and machine learning approaches, incorporating domain-specific fraud indicators, thereby demonstrating the value of combining automated detection with interpretable business rules [26].

2.7. Domain-Specific Applications and Case Studies

Numerous studies have focused on specific healthcare systems and their unique challenges. Kittoe and Asiedu-Addo (2017) employed data mining techniques on Ghana's National Health Insurance Scheme, identifying specific fraud patterns such as excessive prescription of drugs (35% of cases), duplicate registrations (32%), and overbilling (14%) [3]. Alhassan et al. (2016) conducted a comprehensive review of Ghana's NHIS, highlighting persistent challenges, including provider payment delays, fraud and abuse, and governance inefficiencies [27]. Sun et al. (2024) examined fraudulent reimbursement patterns using ARMA modeling for trend forecasting and risk governance [28]. Mailloux et al. (2010) developed decision support tools for identifying controlled substance abuse among Medicaid members using CHAID decision trees, demonstrating the potential of statistical modeling in automating abuse detection while aligning with manual pharmacist reviews [29].

2.8. Emerging Technologies and Future Directions

Recent research has explored the integration of emerging technologies in fraud detection. Kapadiya et al. (2022) proposed a hybrid architecture combining AI and blockchain technology, utilizing machine learning for pattern recognition and blockchain for secure, transparent, and tamper-proof data management [30]. This approach addresses the traditional limitations of reactive, siloed fraud detection methods and provides a foundation for proactive, intelligent fraud prevention systems. Jillo (2024) conducted a comprehensive review of advances and challenges in medical insurance fraud detection, examining both traditional approaches and advanced methods including machine learning, AI, and data mining. The study highlighted the effectiveness of integrating advanced analytics, blockchain, and AI in improving detection accuracy and reducing fraudulent payouts, while identifying persistent challenges including data privacy concerns, high false-positive rates, and complex regulatory compliance [31]. A critical aspect highlighted in multiple studies is the challenge of evaluating fraud detection models under extreme class imbalance. Herland et al. (2019) systematically examined how varying levels of class rarity impact model evaluation, advocating for cost-sensitive metrics and proper cross-validation strategies [32]. This study emphasizes that commonly used metrics, such as AUC, can be misleading under extreme class imbalances, necessitating careful consideration of evaluation methodologies in fraud detection research. Despite the significant progress, several research gaps remain evident. The lack of standardized evaluation protocols, limited availability of labeled datasets, and insufficient attention to evolving fraud patterns represent ongoing challenges. Additionally, while explainability has gained attention, there remains a need for more sophisticated interpretability techniques that can provide actionable insights for fraud investigators. The integration of real-time detection capabilities, adaptive learning mechanisms, and cross-domain knowledge transfer also presents promising research directions for future work in healthcare fraud detection systems.

Table 1 provides a comprehensive summary of the various approaches to fraud detection in insurance and healthcare systems.

Table 1. Summary of Machine Learning Approaches for Fraud Detection in Insurance and Healthcare Systems

Study	**Domain**	**Methods**	**Dataset**	**Key Performance**	**Main Contributions**
Varmedja et al. (2019)[1]	Credit Card Fraud	LR, RF, NB, MLP + SMOTE	Kaggle Credit Card Dataset (284,807 transactions, 0.173% fraud)	RF: 96.38% precision, 81.63% recall, 99.96% accuracy	Demonstrated RF superiority in credit card fraud; effective SMOTE application

Continued on next page

Table 1 – continued from previous page

Study	Domain	Methods	Dataset	Key Performance	Main Contributions
Nabrawi & Alanazi (2023)[5]	Healthcare Insurance	Random Forest	Healthcare insurance claims	Not specified	Identified demographic patterns (42% fraud in ages 23-45); RF effectiveness in healthcare
Sumalatha & Prabha (2019) [9]	Mediclaim	Logistic Regression + MCDA	Mediclaim data	Reasonable accuracy	Integration of predictive analytics with multi-criteria decision analysis
Severino & Peng (2021) [6]	Property Insurance	LR, RF, GBM, XGBoost, LightGBM + SMOTE/Undersampling	Motor vehicle and home insurance	Ensemble methods achieved higher AUPRC	Comparative analysis showing ensemble method superiority in property insurance
Hancock et al. (2023) [2]	Medicare Fraud	LR, RF, XGBoost, LightGBM, CatBoost, ET	CMS Part B & D data	XGBoost: 0.9408 AUPRC	Explainable ML for Medicare; feature selection improving interpretability
Kittoe & Asiedu-Addo (2017) [3]	Health Insurance (NHIS Ghana)	Data Mining Techniques	720 malaria cases (2013)	35% excessive prescription, 32% duplicate registrations	Identified specific fraud patterns in developing country context
Zhang et al. (2020) [26]	Medical Fraud	Hybrid Rule-based + ML	Medical records	High precision for specific fraud types	Integration of domain knowledge with automated detection
du Preez et al. (2024) [4]	Healthcare Claims (Systematic Review)	Various ML methods reviewed	Multiple healthcare datasets	F1-scores up to 0.97, AUPRC ∼0.94	Comprehensive review identifying trends and standardization gaps
Wang et al. (2025) [11]	Healthcare Insurance	Ensemble (XGBoost, RF, LR) + SHAP	Healthcare insurance claims	High AUPRC and F1-scores	Robust ensemble with explainability through SHAP integration
Zhou et al. (2023) [18]	Health Insurance (Collusive Fraud)	Visual Analytics, Network Analysis, Community Detection	Health insurance claims	Successfully identified fraud rings	Novel visual analytics approach for detecting coordinated fraud

Continued on next page

Table 1 – continued from previous page

Study	Domain	Methods	Dataset	Key Performance	Main Contributions
Yoo et al. (2023) [19]	Medicare Fraud	Traditional ML vs GNNs (Graph-SAGE, GGNN, GAT)	CMS Medicare Part B & D	CatBoost: 0.91 AUC-ROC, GraphSAGE: 0.87 AUC-ROC	Comparative study of traditional ML vs graph-based approaches
Kapadiya et al. (2022) [30]	Healthcare Insurance	AI + Blockchain Architecture	Conceptual framework	Improved accuracy, reduced delays	Novel integration of AI and blockchain for fraud detection
Sun et al. (2024) [28]	Healthcare Reimburse-ment	ARMA Model	2018-2021 health insurance data	Trend forecasting capability	Proactive risk governance through temporal modeling
Alhassan et al. (2016) [27]	Health Insurance (NHIS Ghana)	Policy Review/Analysis	NHIS Ghana operational data	Identified systemic issues	Comprehensive policy analysis identifying fraud vulnerabilities

3. Materials and Methods

3.1. Dataset Description

This study employed a comprehensive healthcare fraud detection dataset collected from the National Health Insurance Scheme (NHIS) [29], comprising 20,388 medical claim records that included eight essential features that encapsulate essential data regarding patient interactions, billing processes, and fraud patterns. The dataset demonstrates the complexity of real-world healthcare claims, covering various data types, temporal patterns, and multiple fraud categories. Table 2 presents the feature names associated with their data types and descriptions with value ranges or examples of feature values.

As shown in Table 3, The NHIS dataset suggests that 57.41% of cases are legitimate (No Fraud), suggesting a relatively balanced distribution compared to standard fraud detection scenarios, in which fraudulent cases are frequently rare (typically $< 1\%$). This indicates that the dataset may have been intentionally curated or balanced to achieve the research objectives. Two primary types of fraudulent activities were identified: Phantom Billing (20.76%), which entails billing for services that were never provided or for patients that do not exist, and Ghost Enrollee (20.10%), which pertains to the fraudulent enrolment of individuals who are either non-existent or not eligible for services. Wrong Diagnosis fraud constitutes only 1.73% of cases, making it the least prevalent category within this dataset. This may suggest that wrong diagnosis fraud is either infrequent or challenging to detect and document.

3.2. Proposed Framework Architecture

As illustrated in Figure 1, the framework for multi-class healthcare fraud detection offers a systematic strategy that progresses through four interrelated stages, each of which aims to address particular healthcare fraud detection challenges while guaranteeing computational effectiveness and predictive accuracy. By shifting away from traditional binary classification techniques, this integrated methodology enables the thorough identification of fraud categories that enable focused intervention strategies in healthcare systems. The NHIS dataset, which consists of 20,388 medical claims classified by eight unique attributes that form the foundation for additional

Table 2. National Health Insurance Scheme (NHIS) Dataset Feature Descriptions

Feature Name	**Data Type**	**Description**	**Value Range/Examples**
Patient ID	Integer	Unique patient identifier	1–20,388
AGE	Float	Patient age in years	0.0–120.0
Amount Billed		Total claimed amount	$0.00–$50,000+
DATE OF ENCOUNTER	Date	Patient-provider interaction timestamp	YYYY-MM-DD format
DATE OF DISCHARGE		Service completion timestamp	
GENDER	String	Patient gender	M, F
DIAGNOSIS		Medical diagnosis codes	ICD codes, medical terms
FRAUD_TYPE		Multi-class target variable	No Fraud, Phantom Billing, Wrong Diagnosis, Ghost Enrollee

Table 3. Healthcare Fraud Classification Distribution

Class ID	**Class Name**	**Ratio (%)**
1	No Fraud	57.41
2	Phantom Billing	20.76
3	Ghost Enrollee	20.10
4	Wrong Diagnosis	1.73

analytical processes, is where the framework starts. Essential transformations that handle the diverse nature of healthcare data, like temporal feature conversion and thorough approaches for assigning missing values, are part of the data preprocessing phase. The preprocessing operations maintain data consistency and compatibility with subsequent neural network architectures while preserving the underlying patterns of the original feature space. After preprocessing, the feature engineering phase employs advanced encoding techniques and normalization methods to convert categorical variables into numerical formats that are appropriate for neural network analysis. This phase preserves the semantic integrity of healthcare-specific variables and optimizes their mathematical properties for gradient-based optimization algorithms. The normalization procedures guarantee that features across various scales contribute fairly to the learning process, thereby preventing any individual feature from dominating the model training dynamics. To identify the most informative features while reducing dimensionality and preventing overfitting, we employed an ensemble-based voting feature selection approach combining four complementary feature selection methods:

1. **Variance Threshold** (threshold = 0.01): Eliminates features with near-zero variance, removing constants or quasi-constants that provide no discriminative information.

2. **Mutual Information** (SelectKBest, k=15): Measures dependency between each feature and the target variable, capturing both linear and nonlinear relationships:

$$MI(X;Y) = \sum_{y\in Y}\sum_{x\in X} p(x,y)\log\frac{p(x,y)}{p(x)p(x)} \tag{1}$$

3. **Tree-Based Importance** (Random Forest, n_estimators=100): Computes Gini importance scores reflecting the total reduction in node impurity weighted by the probability of reaching each node.

4. **Recursive Feature Elimination** (RFE with Random Forest estimator): Iteratively removes least important features, reassessing importance at each step to capture feature interaction effects.

Ensemble Voting Procedure: Each method ranks features by importance. A feature is retained if it appears in the top 50% for at least 3 out of 5 methods. This consensus approach balances multiple criteria and reduces the risk of overfitting to any single selection heuristic.

Table 4. Top 8 Feature Importance Scores from ensemble-based voting feature selection

Feature	Variance	MI Score	RF Importance	Ensemble Votes
Amount Billed	1.00	0.339	0.228	4
NEGATIVE_STAY	1.00	0.143	0.075	4
DISCHARGE_DAY_OF_WEEK	1.00	0.109	0.058	4
AGE_GROUP_FREQ_0.37	0.233	0.060	0.051	4
ENCOUNTER_DAY_OF_WEEK	1.00	0.106	0.045	4
AGE	1.00	0.160	0.042	4
ENCOUNTER_MONTH	1.00	0.127	0.028	4
AGE_GROUP_46-65	0.233	0.060	0.029	4

The ensemble approach identified Amount Billed, NEGATIVE_STAY, and DISCHARGE_DAY_OF_WEEK as the most discriminative features, consistent with domain expertise in healthcare fraud investigation where billing anomalies and demographic risk factors are primary indicator. The architectural core of the framework comprises four distinct deep neural network models, each designed to capture complementary aspects of healthcare fraud patterns using various representational strategies. The Simple Neural Network functions as an efficient baseline, utilizing a streamlined architecture that balances predictive performance and deployment practicality. The Deep Wide Neural Network enhances the representational capacity by increasing both the depth and width, facilitating the identification of complex nonlinear relationships in high-dimensional fraud patterns. The Regularized Neural Network effectively mitigates overfitting by employing a combination of L1 and L2 penalty mechanisms, along with strategic dropout placement, thereby promoting robust generalization in various healthcare settings. The Residual Neural Network employs skip connections to enhance representation learning and reduce gradient degradation problems typically found in conventional feedforward architectures. The framework results in a multi-class prediction system that categorizes healthcare claims into four distinct fraud typologies: No Fraud for legitimate claims, Phantom Billing for charges related to non-provided services, Wrong Diagnosis for intentional diagnostic misclassification aimed at increased reimbursement, and Ghost Enrollee for fraudulent use of fictitious patient identities.

3.3. Deep Neural Network Architectures

This study presents four unique deep neural network designs, each carefully designed to tackle the complexity of multi-class healthcare fraud detection while identifying varied patterns in healthcare claims data. Architectural diversity facilitates the extensive evaluation of several neural network architectures and their relevance to fraud detection in the healthcare sector.

3.3.1. Simple Neural Network The Simple Neural Network acts as the baseline model, employing a feedforward architecture that successfully balances computational efficiency and prediction accuracy. This architecture systematically reduces the neuronal density across three hidden layers, starting with 64 neurons in the first layer, 32 in the second, and ending with 16 neurons in the third layer. Each hidden layer employs batch normalization to stabilize the training behavior and reduce the internal covariate shift, whereas dropout regularization at a rate of 0.3 is utilized to prevent overfitting. The model accepts pre-processed input features, with dimensionality established via systematic feature selection methods. The architectural design adheres to the principle of progressive

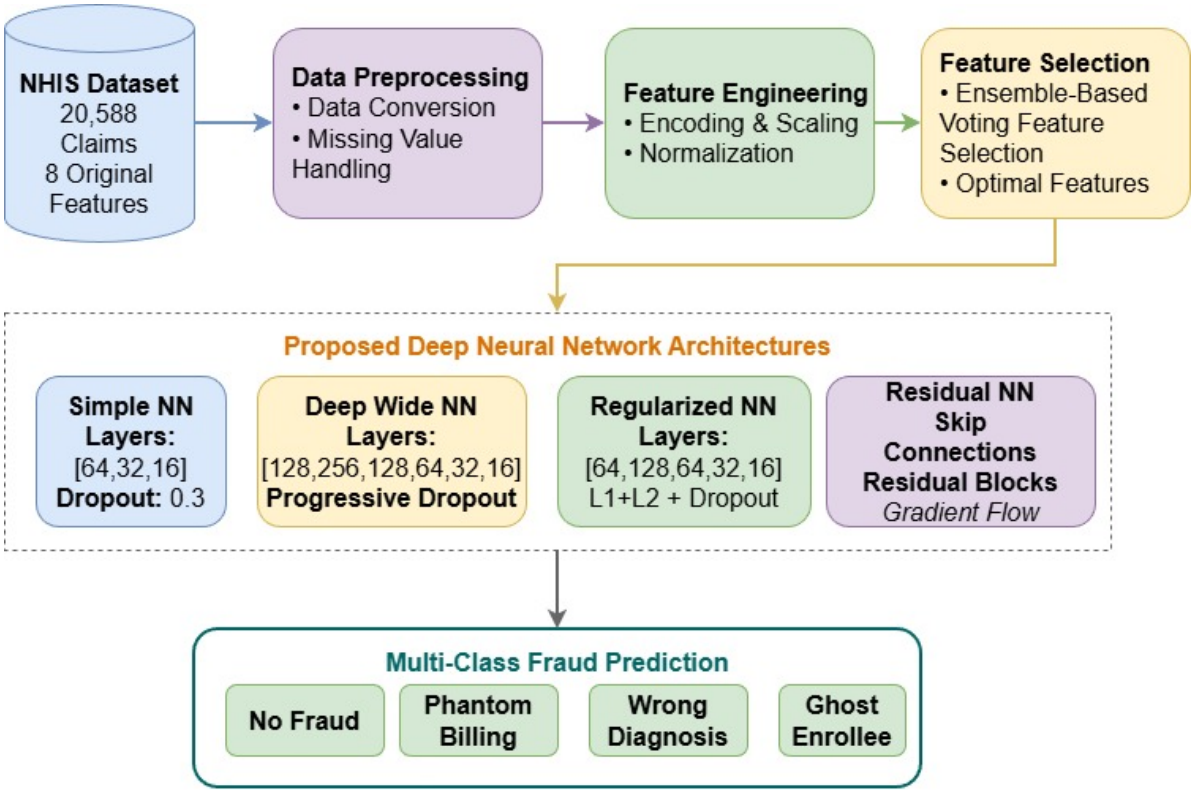

Figure 1. Pipeline Architecture for the proposed Multi-Class Healthcare Fraud

dimensionality reduction, which enables the network to acquire hierarchical representations of fraud patterns. The final layer utilizes softmax activation to generate probability distributions among the four fraud categories: No Fraud, Phantom Billing, Wrong Diagnosis, and Ghost Enrollee. The Simple Neural Network aims to learn a mapping function $f : \mathcal{X} \to \mathcal{Y}$ that minimizes the categorical cross-entropy loss between the true labels $\mathbf{y}$ and predictions $\hat{\mathbf{y}}$:

$$\mathcal{L}_{\text{simple}}(\theta) = -\sum_{i=1}^{N}\sum_{j=1}^{K} y_{ij}\log(\hat{y}_{ij}) \tag{2}$$

where:

- N is the number of samples
- K is the number of classes ($K = 4$ in our multi-class problem)
- y_{ij} is the true label (one-hot encoded) for sample i and class j
- $\hat{y}_{ij} = \text{softmax}(f(\mathbf{x}_i;\theta))_j$ is the predicted probability for class j
- θ represents all trainable parameters (weights and biases)

The network employs the Adam optimizer to minimize this objective through stochastic gradient descent with momentum and adaptive learning rates. The learning process iteratively updates parameters via:

$$\theta_{t+1} = \theta_t - \alpha \cdot \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} \tag{3}$$

where $\hat{m}_t$ and $\hat{v}_t$ are bias-corrected first and second moment estimates, $\alpha = 0.001$ is the learning rate, and $\epsilon = 10^{-8}$ is a small constant for numerical stability.

The model utilizes the Adam optimizer with a learning rate of 0.001 and employs categorical crossentropy as the loss function to address the multi-class classification objective. The architecture comprises approximately 100,324 trainable parameters, indicating the most efficient design among the proposed models.

3.3.2. Deep Wide Neural Network The Deep Wide Neural Network significantly increases both depth and width relative to the baseline model, aiming to capture complex nonlinear relationships and intricate patterns in healthcare fraud data through improved representational capacity. This architecture utilizes six hidden layers characterized by a unique expansion-contraction pattern: the network begins with 128 neurons in the first layer, expands to 256 neurons in the second layer, and subsequently contracts through 128, 64, 32, and 16 neurons in the following layers. The expanded architecture requires advanced regularization strategies to preserve its generalization ability. The Deep Wide Neural Network optimizes the same categorical cross-entropy objective as the Simple NN but with enhanced representational capacity through increased depth and width:

$$\mathcal{L}_{\text{wide}}(\theta) = -\sum_{i=1}^{N}\sum_{j=1}^{K} y_{ij}\log(\hat{y}_{ij}) \tag{4}$$

The increased architectural complexity enables the learning of hierarchical representations through compositional mappings:

$$f(\mathbf{x};\theta) = f_L \circ f_{L-1} \circ \cdots \circ f_2 \circ f_1(\mathbf{x}) \tag{5}$$

where each layer $l \in \{1, 2, \ldots, L\}$ applies the transformation:

$$\mathbf{h}_l = \text{Dropout}(\text{ReLU}(\text{BatchNorm}(\mathbf{W}_l\mathbf{h}_{l-1} + \mathbf{b}_l)); p_l) \tag{6}$$

Here, $\mathbf{W}_l$ and $\mathbf{b}_l$ denote the weight matrix and bias vector for layer l, and p_l represents the layer-specific dropout probability. The progressive dropout strategy ($p_{\text{early}} = 0.4 \rightarrow p_{\text{late}} = 0.2$) provides implicit regularization without explicit penalty terms, allowing the expanded architecture to maintain generalization while capturing intricate fraud patterns across multiple levels of abstraction. The model employs progressive dropout rates, starting with 40% in the initial layers, where the risk of overfitting is greatest, and gradually decreasing to 20% in the subsequent layers. This graduated regularization method recognizes the different complexities of representations acquired at various depths within the network. The optimizer utilizes a learning rate of 0.0005 to manage the expanded parameter space and promote stable convergence throughout the training process. The model comprises approximately 500,000 trainable parameters, indicating a five-fold increase in complexity relative to the baseline architecture.

3.3.3. Regularized Neural Network The Regularized Neural Network uses explicit regularization techniques to mitigate the issues of overfitting and inadequate generalization frequently observed in healthcare fraud detection, which arise from class imbalance and high-dimensional feature spaces. This architecture employs combined L1 and L2 regularization, utilizing coefficients λ_1 = 0.01 and λ_2 = 0.01, uniformly across all hidden layers. The network topology comprises five hidden layers with dimensions [64, 128, 64, 32, 16], establishing a symmetric expansion-contraction pattern that enhances feature extraction and dimensionality reduction. The regularization strategy encompasses not only weight penalties but also significant dropout rates, transitioning from 50% in the initial layers to 30% in the final hidden layers. This dual regularization method guarantees strong generalization while preserving the adequate model capacity to capture intricate fraud patterns.

The Regularized Neural Network explicitly incorporates sparsity-inducing and weight magnitude constraints into its objective function to combat overfitting in high-dimensional healthcare fraud data. The model optimizes a composite loss function that balances predictive accuracy with parameter complexity:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{ce}} + \lambda_1 \sum_i |w_i| + \lambda_2 \sum_i w_i^2 \tag{7}$$

where:

- $\mathcal{L}_{\text{ce}}$ denotes the categorical cross-entropy loss

- $\lambda_1 = 0.01$ is the L1 regularization coefficient, promoting sparsity
- $\lambda_2 = 0.01$ is the L2 regularization coefficient, controlling weight magnitude
- w_i indicates individual network weights across all layers

The L1 penalty encourages sparse solutions by driving less important weights toward zero:

$$\frac{\partial}{\partial w_i}\left(\lambda_1 |w_i|\right) = \lambda_1 \cdot \text{sign}(w_i) \tag{8}$$

while the L2 penalty prevents any single weight from dominating the model output:

$$\frac{\partial}{\partial w_i}\left(\lambda_2 w_i^2\right) = 2\lambda_2 w_i \tag{9}$$

This dual regularization approach, combined with aggressive dropout rates (0.5 in early layers, 0.3 in later layers), ensures robust generalization by constraining the hypothesis space while preserving adequate model capacity to capture complex fraud patterns. The symmetric expansion-contraction architecture $[64, 128, 64, 32, 16]$ further facilitates feature extraction and dimensionality reduction in a principled manner.

The total loss function integrates the classification loss with regularization penaltieFs as follows:

$$L_{\text{total}} = L_{\text{ce}} + \lambda_1 \sum |w_i| + \lambda_2 \sum w_i^2 \tag{10}$$

Where:

- L_{ce} denotes the categorical cross-entropy loss
- λ_1 and λ_2 signify the L1 and L2 regularization coefficients respectively
- w_i indicates individual network weights

This formulation enhances sparsity via L1 regularization and controls the weight magnitude through L2 regularization, leading to a model that preserves predictive accuracy and demonstrates improved generalization capabilities.

3.3.4. Residual Neural Network The Residual Neural Network applies the creative idea of skip connections, initially created for deep convolutional networks in computer vision, to the field of tabular healthcare data. This adaptation mitigates the vanishing gradient problem, which often limits the training of deeper feedforward networks, facilitating the improved learning of complicated fraud patterns via enhanced information flow.

The Residual Neural Network optimizes the standard categorical cross-entropy loss while employing skip connections to facilitate gradient flow and enable the training of deeper architectures:

$$\mathcal{L}_{\text{residual}}(\theta) = -\sum_{i=1}^{N}\sum_{j=1}^{K} y_{ij} \log(\hat{y}_{ij}) \tag{11}$$

Unlike conventional feedforward networks, each residual block learns a residual mapping rather than a direct transformation. For a residual block with input $\mathbf{x}$ and output $\mathbf{y}$, the relationship is defined as:

$$\mathbf{y} = \mathcal{F}(\mathbf{x}, \{\mathbf{W}_i\}) + \mathbf{x} \tag{12}$$

where $\mathcal{F}(\mathbf{x}, \{\mathbf{W}_i\})$ represents the residual function to be learned (typically comprising two or more weighted layers with nonlinear activations), and the identity mapping $\mathbf{x}$ is added via the skip connection. This formulation allows the network to learn incremental transformations, with the optimization focusing on residual adjustments $\mathcal{F}(\mathbf{x})$ rather than complete transformations $\mathbf{y} = \mathcal{H}(\mathbf{x})$.

The skip connections fundamentally alter the gradient propagation dynamics. During backpropagation, the gradient with respect to the input includes a direct path through the identity connection:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{x}} = \frac{\partial \mathcal{L}}{\partial \mathbf{y}} \cdot \left(\frac{\partial \mathcal{F}}{\partial \mathbf{x}} + \mathbf{I} \right) \tag{13}$$

where $\mathbf{I}$ is the identity matrix. This ensures stable gradient propagation even in deeper architectures, mitigating the vanishing gradient problem that often limits conventional feedforward networks. The additive nature of skip connections guarantees that gradients can flow directly through the network without being attenuated by multiple sequential nonlinearities, enabling effective training of representations at varying levels of abstraction.

The architecture incorporates two separate residual blocks with aligned dimensional constraints to enable element-wise addition. The initial residual block operates at 64 dimensions, whereas the subsequent block functions at 128 dimensions. Each residual block adheres to the essential formulation y = F(x) + x, where x denotes the block input, F(x) the learned transformation, and y the block output. This formulation facilitates the direct flow of gradient information through skip connections, thereby reducing the gradient degradation in deeper networks. The architecture consists of eight layers, incorporating strategically placed residual connections that maintain the information flow and facilitate the learning of incremental transformations. Skip connections enhance the training of deeper representations, reducing the optimization challenges commonly associated with greater network depth. Batch normalization and dropout regularization were consistently implemented across the architecture to ensure training stability and minimize overfitting.

3.3.5. Hyperparameter Selection and Justification Hyperparameter selection significantly impacts neural network performance, particularly for fraud detection where suboptimal configurations can lead to either underfitting (missing fraud patterns) or overfitting (poor generalization). We employed systematic grid search with 5-fold stratified cross-validation to identify optimal hyperparameters for each architecture.

Table 5. Hyperparameter Search Space and Final Selected Values

Architecture	Hyperparameter	Search Space	Final Value
5*Simple NN	Learning Rate	[0.0001, 0.0005, 0.001, 0.005]	0.001
	Batch Size	[32, 64, 128]	64
	Dropout Rate	[0.2, 0.3, 0.4, 0.5]	0.3
	Hidden Layers	[(32,16), (64,32), (64,32,16)]	(64,32,16)
	Activation	[relu, tanh, elu]	relu
4*Deep Wide NN	Learning Rate	[0.0001, 0.0005, 0.001]	0.0005
	Dropout (Early)	[0.3, 0.4, 0.5]	0.4
	Dropout (Late)	[0.2, 0.3, 0.4]	0.2
	Layer Sizes	Various configurations	[128,256,128,64,32,16]
4*Regularized NN	L1 Penalty (λ_1)	[0.001, 0.01, 0.1]	0.01
	L2 Penalty (λ_2)	[0.001, 0.01, 0.1]	0.01
	Dropout Rate	[0.3, 0.4, 0.5]	0.5 (early), 0.3 (late)
	Learning Rate	[0.0005, 0.001, 0.002]	0.001

Rationale for Key Choices:

1. **Learning Rate (0.0005-0.001):** Selected based on loss convergence curves. Higher rates (> 0.005) caused training instability, while lower rates (< 0.0001) required prohibitively long training times without performance gains.

2. **Batch Size (64):** Balances computational efficiency with gradient estimate quality. Smaller batches (32) increased training time by 40% with minimal accuracy gain, while larger batches (128) degraded generalization by 1-2%.

3. **Dropout Rates (0.2-0.5):** Progressive dropout reflects layer-specific overfitting risk. Early layers capture general patterns requiring stronger regularization, while later layers learn class-specific features benefiting from lower dropout.

4. **Regularization Coefficients** ($\lambda = 0.01$)**:** Grid search revealed that $\lambda < 0.001$ provided insufficient regularization (train-test gap $> 5\%$), while $\lambda > 0.1$ caused underfitting (validation accuracy $< 75\%$). $\lambda = 0.01$ achieved optimal bias-variance balance.

This systematic exploration required training 148 model configurations, using early stopping with patience=20 epochs to prevent overfitting during hyperparameter search.

Table 6 summarizes the four proposed deep neural network architectures, highlighting their distinctive features and design principles.

Table 6. Comparative Analysis of Proposed Deep Neural Network Architectures

	Simple NN	**Deep Wide NN**	**Regularized NN**	**Residual NN**
Architecture Type	Feedforward	Deep Wide Network	Explicit Regularization	Skip Connection Network
Hidden Layers	3	6	5	8 (with residual blocks)
Layer Dimensions	[64, 32, 16]	[128, 256, 128, 64, 32, 16]	[64, 128, 64, 32, 16]	[64, 64, 128, 128, 64, 32, 16]
Trainable Parameters	∼100,324	∼500,000+	∼200,000	∼300,000
Regularization Strategy	Dropout (0.3)	Progressive Dropout (0.4→0.2)	L1+L2 + Dropout (0.5→0.3)	Standard Dropout (0.3→0.2)
L1 Regularization	None	None	$\lambda_1 = 0.01$	None
L2 Regularization	None	None	$\lambda_2 = 0.01$	None
Skip Connections	None	None	None	2 Residual Blocks
Batch Normalization	All Hidden Layers	All Hidden Layers	All Hidden Layers	All Hidden Layers
Learning Rate	0.001	0.0005	0.001	0.001
Optimizer	Adam	Adam	Adam	Adam

The architectural analysis demonstrates design principles that address various facets of healthcare fraud detection. The Simple Neural Network emphasizes computational efficiency and interpretability, rendering it suitable for resource-constrained environments and real-time deployment scenarios. The Deep Wide Neural Network enhances representational capacity to identify complex fraud patterns, although increased computational demands and a risk of overfitting. The Regularized Neural Network effectively tackles generalization issues by employing robust regularization techniques, making it especially appropriate for situations characterized by constrained training data or higher noise levels. The Residual Neural Network facilitates enhanced representation learning by preserving gradient flow, providing an advanced method for complicated pattern recognition.

4. Results and Discussion

4.1. Training Configuration and Evaluation Metrics

The evaluation of the proposed healthcare fraud detection system used a comprehensive experimental framework that include four distinct deep learning architectures: Simple NN, Deep Wide NN, Regularized NN, and Residual NN. The dataset of 20,388 healthcare claims was partitioned into training, validation, and test sets, dividing 2,039 samples for final testing to four fraud categories: Ghost Enrollee (410 samples), No Fraud (1,171 samples), Phantom Billing (423 samples), and Wrong Diagnosis (35 samples). The assessment of model performance utilized various evaluation metrics, including accuracy, precision, recall, and F1-scores, that includes both macro and weighted averages, equations (1 to 10). Training durations varied from 40.6 seconds for the Simple Neural Network

to 156.8 seconds for the Regularized Neural Network, indicating computational efficiency across all models. The models exhibited considerable variation in complexity, with parameter counts ranging from 100,324 for the Simple NN to 489,188 for the Deep Wide NN.

$$\text{Accuracy} = \frac{\sum_{i=1}^{k} TP_i}{N} \tag{14}$$

$$\text{Precision}_i = \frac{TP_i}{TP_i + FP_i} \tag{15}$$

$$\text{Recall}_i = \frac{TP_i}{TP_i + FN_i} \tag{16}$$

$$F1_i = 2 \times \frac{\text{Precision}_i \times \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i} \tag{17}$$

$$\text{Precision}_{\text{macro}} = \frac{1}{k} \times \sum_{i=1}^{k} \text{Precision}_i \tag{18}$$

$$\text{Recall}_{\text{macro}} = \frac{1}{k} \times \sum_{i=1}^{k} \text{Recall}_i \tag{19}$$

$$F1_{\text{macro}} = \frac{1}{k} \times \sum_{i=1}^{k} F1_i \tag{20}$$

$$\text{Precision}_{\text{weighted}} = \frac{\sum_{i=1}^{k} n_i \times \text{Precision}_i}{N} \tag{21}$$

$$\text{Recall}_{\text{weighted}} = \frac{\sum_{i=1}^{k} n_i \times \text{Recall}_i}{N} \tag{22}$$

$$\text{F1}_{\text{weighted}} = \frac{\sum_{i=1}^{k} n_i \times F1_i}{N} \tag{23}$$

where:

- TP_i = True Positives for class i
- FP_i = False Positives for class i
- FN_i = False Negatives for class i
- n_i = Number of samples in class i (support)
- N = Total number of samples
- $C = \{1, 2, \ldots, k\}$

4.2. Overall Model Performance Analysis and benchmarking

Figure 2 presents a comprehensive three-dimensional comparison of all evaluated models across test accuracy, test AUC, and training time, providing essential insights for model selection in healthcare fraud detection systems. The test accuracy scores ranging from approximately 77.02% to 78.3%. Simple NN emerges as the top performer at approximately 78.02% test accuracy and Test AUC 0.9194. The relatively tight clustering of accuracy scores suggests that all models have successfully learned meaningful fraud detection patterns from the dataset, with the modest performance differences indicating that algorithmic choice may be less critical than other factors such as feature engineering quality. The AUC comparison presents a striking finding in the middle panel, where all six models demonstrate exceptionally strong and nearly identical discriminative ability, with scores clustering tightly

around the mean of 0.917. Every model achieves AUC greater than or equal to 0.91, indicating excellent class separation capability regardless of algorithmic approach or architectural complexity. The training time comparison in the right panel reveals dramatic efficiency differences with profound practical implications, exhibiting a striking thirty-fold range in training duration across models. Random Forest completes training in under five seconds, barely visible on the chart scale, representing extraordinary computational efficiency. XGBoost requires approximately three to four seconds, maintaining minimal computational overhead despite its sophisticated boosting mechanisms. In contrast, Simple NN demands approximately 37 seconds for training, representing an eight to ten-fold increase over baseline models. The computationally intensive deep learning architectures demonstrate even greater overhead, with Deep Wide NN requiring approximately 62 seconds, Residual NN demanding 57 seconds, and Regularized NN exhibiting the longest training time at approximately 135 seconds. These efficiency disparities translate directly to operational constraints in production environments requiring frequent model updates in response to evolving fraud schemes. The visualization provides empirical evidence challenging the assumption that complex deep learning architectures are necessary for healthcare fraud detection at this scale and feature complexity.

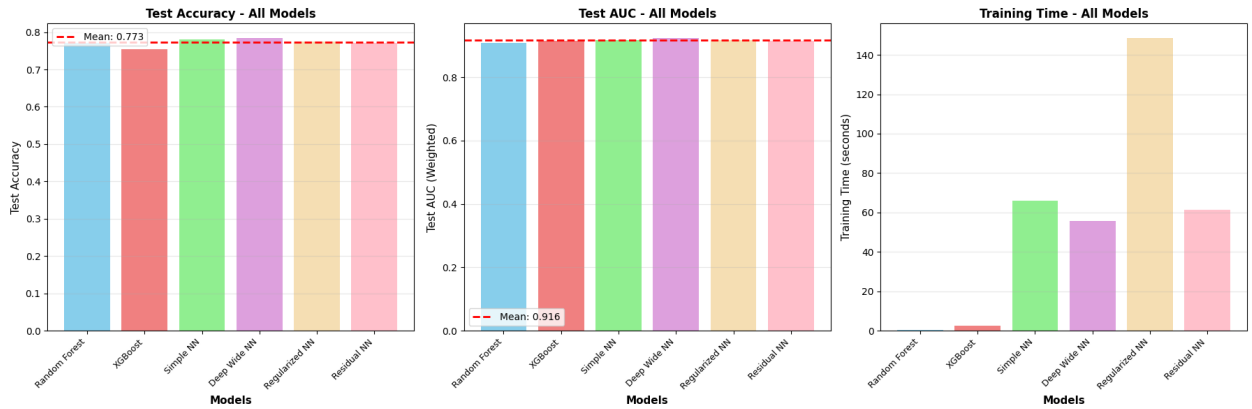

Figure 2. Accuracy Comparison Across Training, Validation, and Test Sets for Multi-Class Models

4.3. Multi-Class Performance Analysis

The analysis of per-class performance indicates considerable variability among fraud types, highlighting the differing complexities involved in detecting various fraudulent behaviors. The confusion matrices shown by figure 3 and per-class performance metrics in figure 4 indicate that Ghost Enrollee fraud exhibits high detectability, with precision at 96.45%, recall at 99.27%, and an F1-score of 97.84%. These results suggest clear and identifiable patterns within this category of fraud.

The No Fraud classification demonstrated strong performance, achieving 88.32% precision and 76.17% recall (F1: 81.80%), which reflects effective identification of legitimate claims alongside acceptable sensitivity levels. The model exhibits a propensity for false positives in fraud detection, as indicated by the precision-recall trade-off illustrated in figure 5.

Phantom Billing displayed the highest difficulty in detection compared to other important fraud categories, achieving a precision of 53.35% and a recall of 71.63% (F1 score: 61.15%). The pattern indicates that the model effectively detects a majority of phantom billing cases; however, it also produces a significant number of false positives. This may suggest the presence of overlapping characteristics with legitimate billing practices or other types of fraud, as evidenced by the confusion matrices shown in Figure 3.

The Wrong Diagnosis fraud, although based on a limited sample size of 35 cases, demonstrated impressive results with a precision of 66.67%, recall of 74.29%, and an F1 score of 70.27%. The limited sample size raises concerns regarding the statistical reliability of these metrics and the model's capacity to generalize to previously unobserved cases in this category.

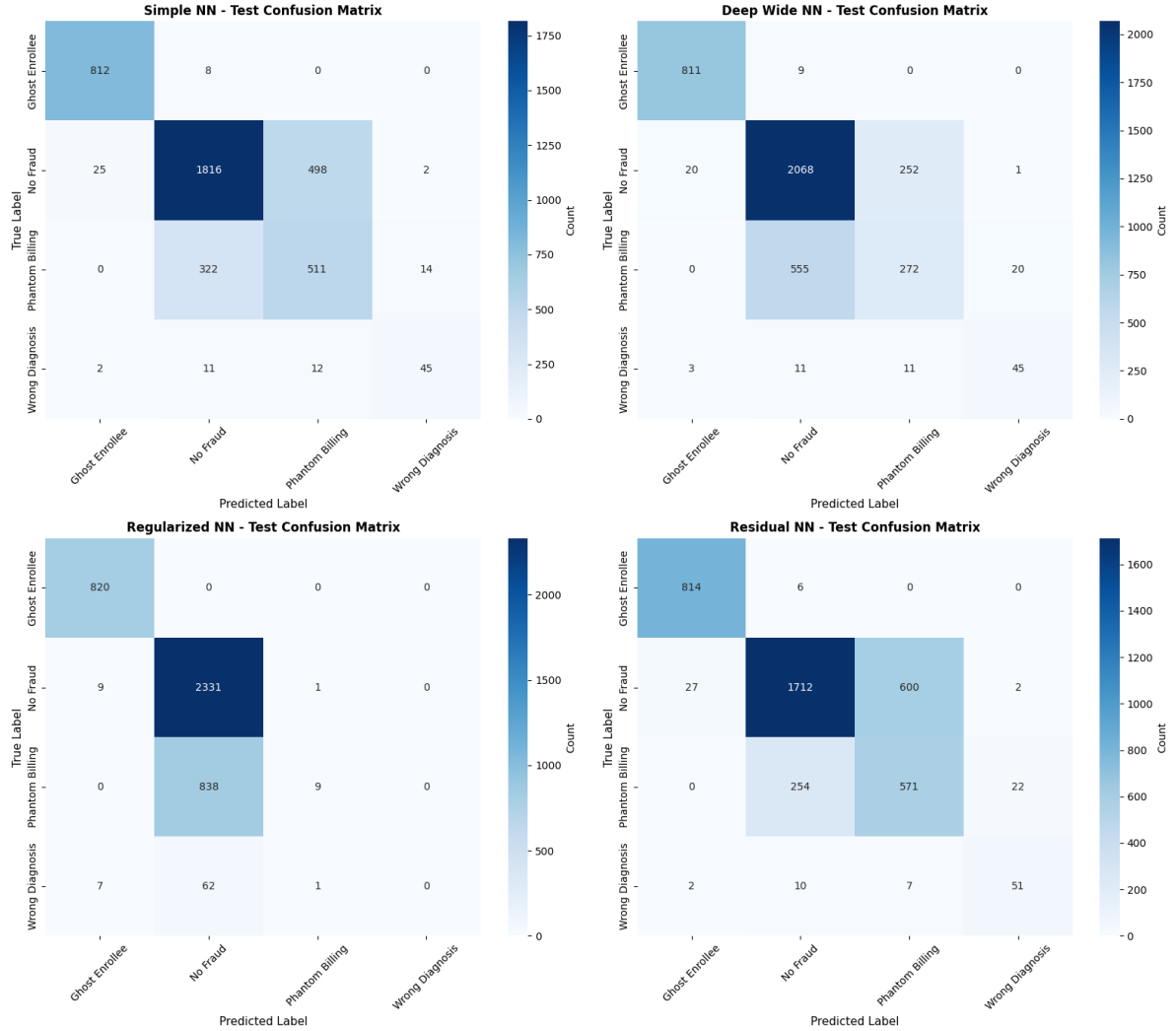

Figure 3. Confusion Matrices for Individual Deep Learning Models (Simple NN, Deep Wide NN, Regularized NN, Residual NN)

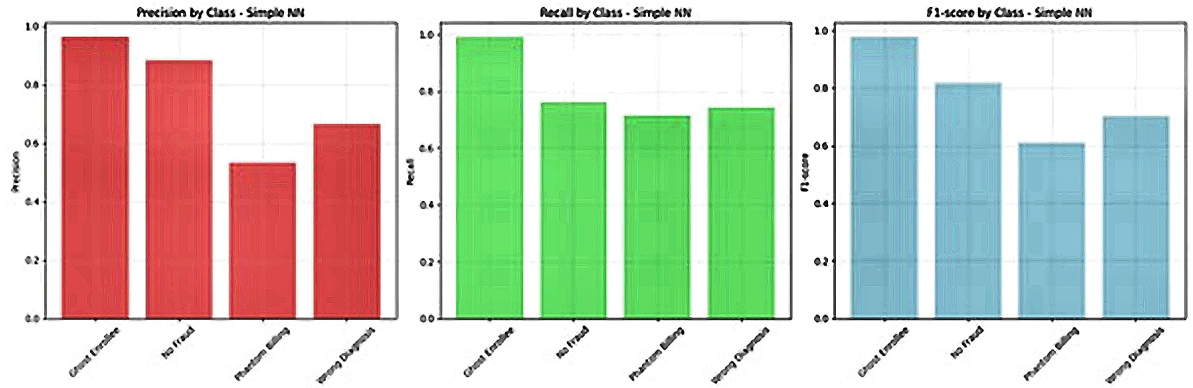

Figure 4. Per-Class Performance Metrics (Precision, Recall, F1-Score) for Simple NN

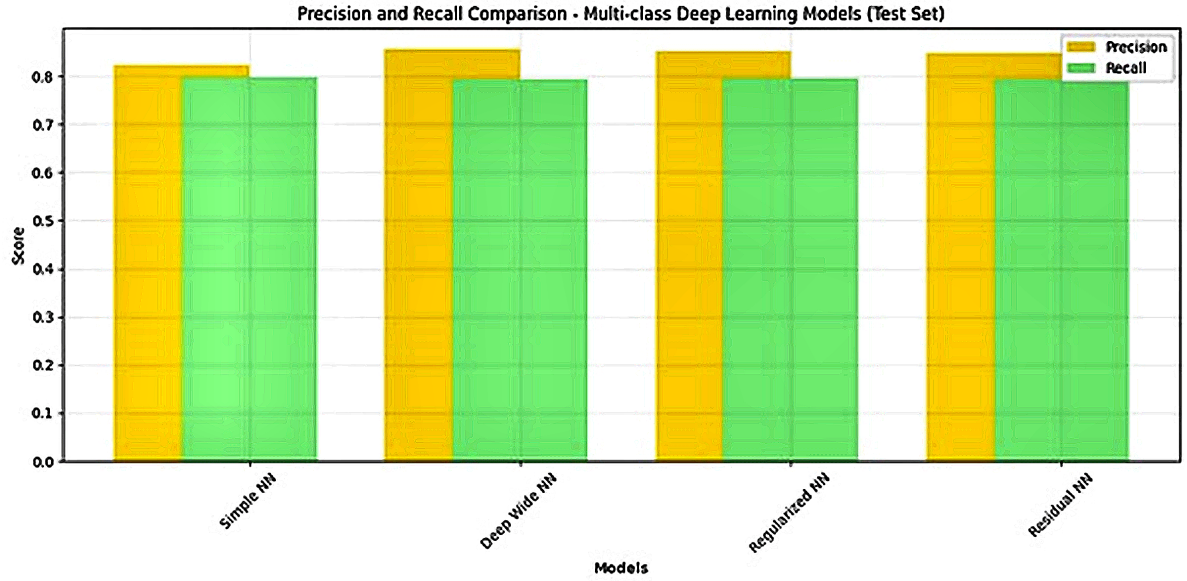

Figure 5. Precision and Recall Performance Comparison for Multi-Class Deep Learning Models

4.4. Overfitting and Generalization Analysis

The overfitting analysis illustrated by Figure 6 demonstrates differing levels of generalization ability among the examined architectures. The Simple NN demonstrates moderate overfitting, evidenced by a train-test accuracy gap of 3.38% (training: 83.22%, test: 79.84%). This suggests a degree of memorization of training patterns while still achieving acceptable generalization.

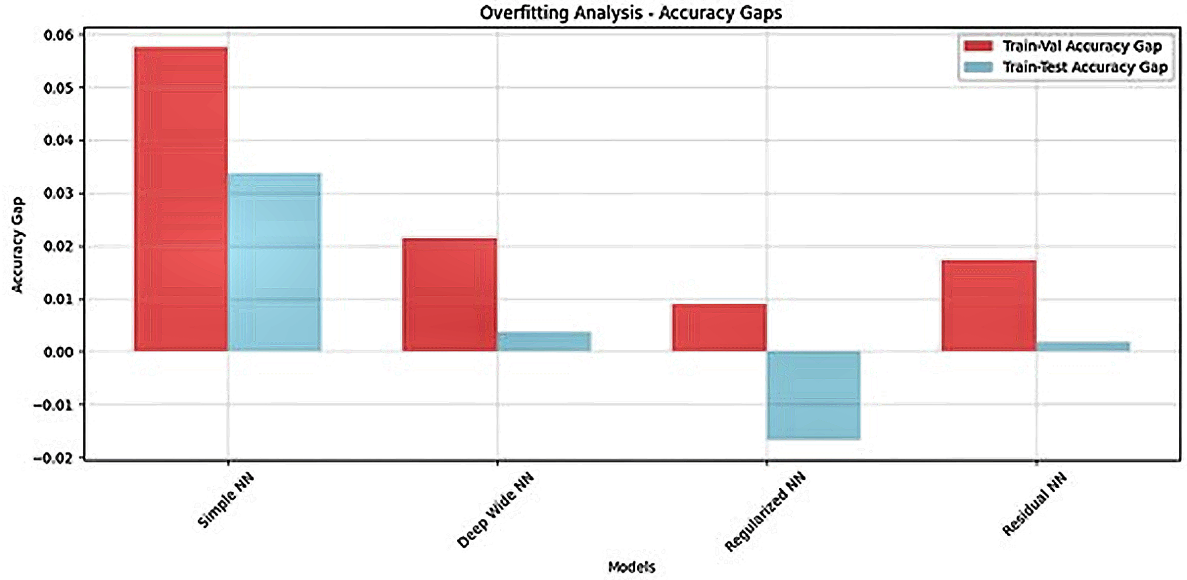

Figure 6. Overfitting Analysis: Train-Validation and Train-Test Accuracy Gaps

The Regularized Neural Network exhibits enhanced generalization, evidenced by a negative train-test gap of -1.68%. This indicates that regularization methods successfully reduced overfitting and may have enhanced test

performance relative to training performance. This unexpected outcome suggests that regularization assisted the model's ability to concentrate on more generalizable patterns. The Deep Wide Neural Network and Residual Neural Network demonstrated strong generalization capabilities, exhibiting minimal discrepancies between training and testing performance (0.36% and 0.17%, respectively). This indicates a natural resistance to overfitting within this domain. The learning curves (Figures 7 and 8) corroborate these findings, indicating stable validation performance across training epochs for the majority of models, while the Simple NN exhibits consistent convergence behavior.

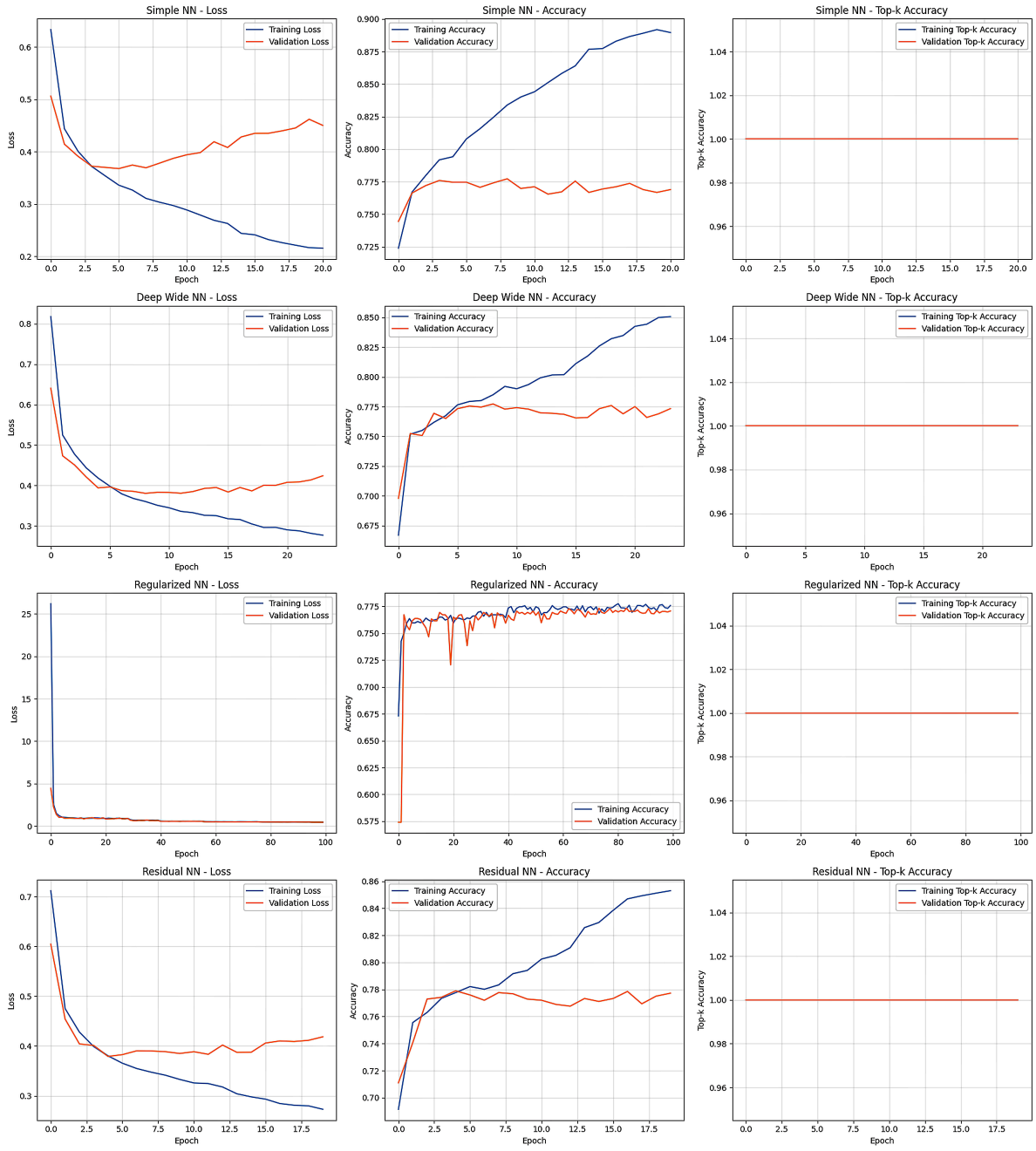

Figure 7. Learning Curves for Multi-Class Deep Learning Models (Loss, Accuracy, and Top-k Accuracy)

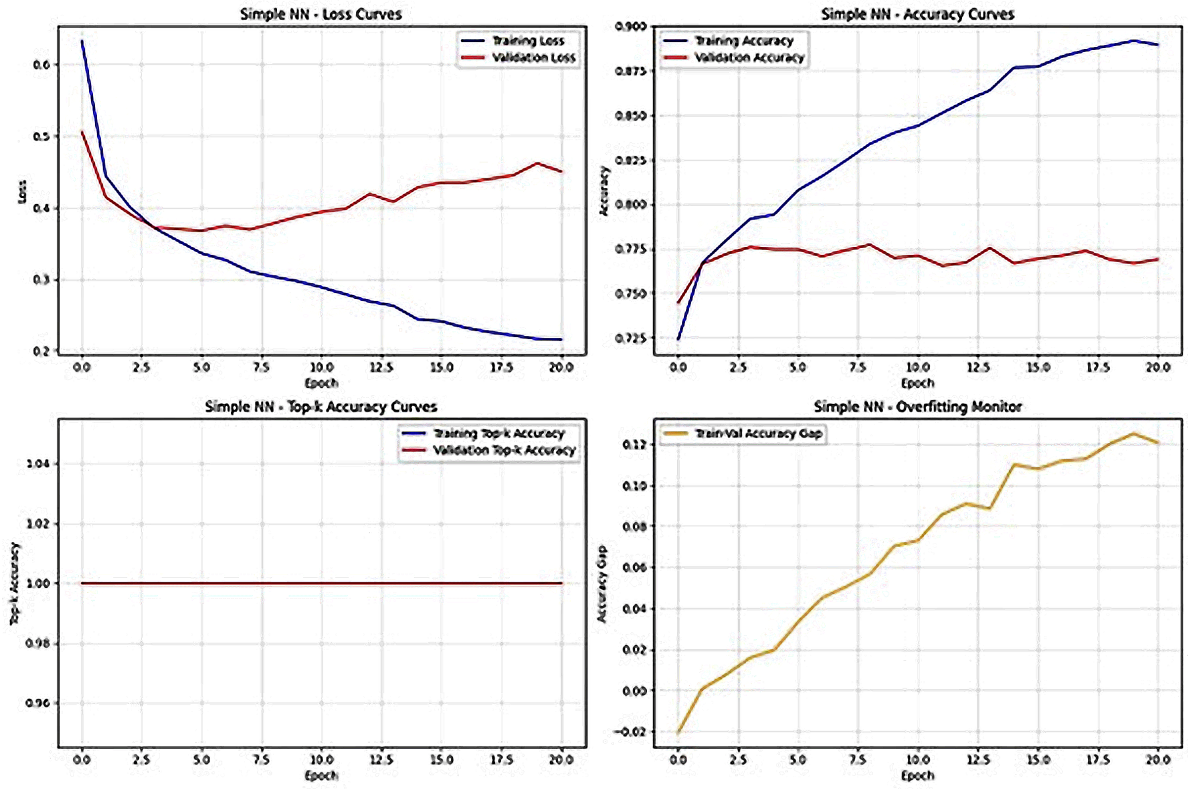

Figure 8. Detailed Learning Curves and Overfitting Monitor for Simple NN Model

The confidence analysis shown by Figure 12 indicates that 57.77% of predictions are classified within the ”Very High” confidence range (0.9-1.0), resulting in an accuracy of 97.20%. The robust correlation between confidence and accuracy offers a practical framework for automated decision-making, enabling the automatic processing of high-confidence predictions while identifying low-confidence cases for manual review.

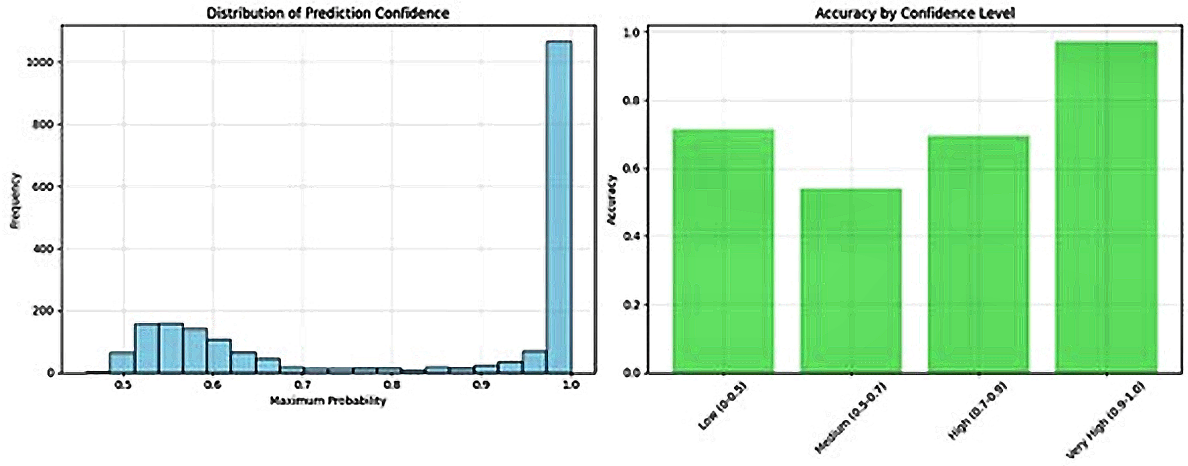

Figure 9. Prediction Confidence Distribution and Accuracy by Confidence Level

The overfitting analysis illustrated by Figure 9 demonstrates differing levels of generalization ability among the examined architectures. The Simple NN demonstrates moderate overfitting, evidenced by a train-test accuracy gap of 3.38% (training: 83.22%, test: 79.84%). This suggests a degree of memorization of training patterns while still achieving acceptable generalization.

4.5. Feature Importance SHAP Analysis

The feature importance analysis using SHAP (SHapley Additive exPlanations) values, as illustrated in Figure 10, reveals critical insights into the Deep Neural Network's decision-making process for healthcare fraud detection. SHAP values quantify the marginal contribution of each feature to model predictions, providing interpretable explanations for the model's classification behavior across the four fraud categories.

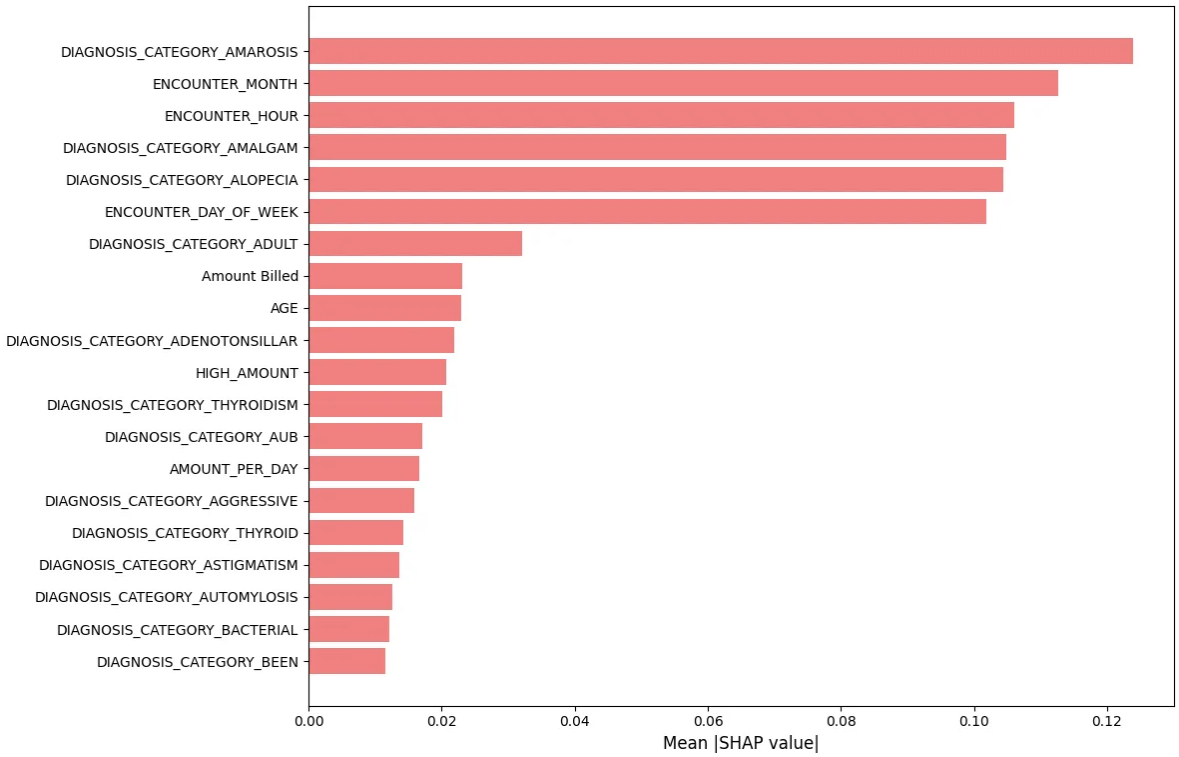

Figure 10. Top 20 Feature Importance Analysis for Deep Neural Network using SHAP Values

The analysis identifies diagnosis category features as the dominant predictors of fraudulent behavior, with DIAGNOSIS_CATEGORY_AMAROSIS emerging as the most influential feature with a mean absolute SHAP value of approximately 0.125. This substantial importance suggests that specific diagnosis codes serve as strong discriminative indicators for fraud detection, potentially reflecting patterns of medical necessity fraud or systematic misuse of particular diagnostic categories. The presence of multiple diagnosis-related features among the top contributors (DIAGNOSIS_CATEGORY_AMALGAM at 0.105, DIAGNOSIS_CATEGORY_ALOPECIA at 0.105, and DIAGNOSIS_CATEGORY_ADULT at 0.032) underscores the central role of diagnostic information in distinguishing fraudulent from legitimate healthcare claims.

Temporal features demonstrate remarkable predictive power, with ENCOUNTER_MONTH (0.115) and ENCOUNTER_HOUR (0.105) ranking as the second and third most important features, respectively. The high importance of ENCOUNTER_MONTH may indicate seasonal patterns in fraudulent activities or systematic billing irregularities that manifest during specific time periods. Similarly, ENCOUNTER_HOUR's prominence suggests that the timing of medical encounters carries significant fraud signals, potentially reflecting unusual consultation

patterns such as excessive after-hours billing or systematic temporal clustering inconsistent with legitimate medical practice. The additional temporal feature ENCOUNTER_DAY_OF_WEEK (0.10) further reinforces the critical role of temporal patterns in fraud detection, indicating that fraudsters may exhibit predictable behavioral patterns across different time dimensions.

Financial features occupy a middle-tier position in the importance hierarchy, with Amount Billed (0.024) and AMOUNT_PER_DAY (0.019) demonstrating moderate but meaningful contributions to fraud prediction. While these features are less influential than diagnosis and temporal indicators, their presence among the top 20 features validates the intuitive expectation that billing amounts serve as fraud indicators. The relatively lower importance of financial features compared to diagnostic and temporal variables suggests that sophisticated fraud schemes may increasingly focus on manipulating diagnostic codes and encounter patterns rather than simply inflating billing amounts, which may be subject to more straightforward automated detection thresholds.

Demographic features, particularly AGE (0.024), show modest but statistically meaningful importance. The age variable's contribution to fraud detection likely reflects age-specific fraud vulnerability patterns or systematic targeting of particular age groups by fraudulent providers. This finding aligns with domain knowledge indicating that certain fraud schemes disproportionately affect specific demographic segments, such as elderly populations being targeted for unnecessary services.

The presence of multiple specific diagnosis categories in the lower tiers of importance (ranging from 0.015 to 0.019) indicates that the model leverages a diverse set of diagnostic indicators beyond the top contributors. Features such as DIAGNOSIS_CATEGORY_THYROIDISM, DIAGNOSIS_CATEGORY_AUB, and various other condition codes collectively contribute to the model's discriminative capability. This distributed importance across multiple diagnosis features suggests that fraudulent patterns manifest across various medical specialties and diagnostic domains, necessitating comprehensive feature coverage rather than reliance on a narrow set of high-importance predictors.

The HIGH_AMOUNT indicator (0.022) appears as a distinct binary flag for elevated billing values, serving as a complementary signal to the continuous Amount Billed feature. Its independent importance suggests that threshold-based billing anomalies provide additional fraud detection value beyond the raw billing amounts themselves.

The feature importance distribution reveals a clear hierarchical structure with three distinct tiers: high-importance features dominated by specific diagnosis categories and temporal variables (SHAP values above 0.10), moderate-importance features including financial and demographic indicators (SHAP values between 0.02 and 0.04), and lower-importance features comprising additional diagnosis categories (SHAP values between 0.015 and 0.02). This hierarchical pattern provides actionable insights for feature engineering priorities in future model iterations and suggests that fraud detection systems should prioritize data quality and completeness for diagnosis codes and temporal information.

The SHAP analysis provides compelling evidence that healthcare fraud detection benefits from a multi-dimensional approach incorporating diagnostic, temporal, financial, and demographic information. The dominance of diagnosis and temporal features challenges traditional fraud detection paradigms that primarily focus on financial anomalies, suggesting that modern deep learning approaches successfully leverage complex behavioral and clinical patterns that may be imperceptible to rule-based systems or simpler statistical methods.

5. Conclusion and Future Directions

This study examines deep learning architectures for multi-class healthcare fraud detection, highlighting the effectiveness of automated classification systems in identifying various fraud patterns. The Simple Neural Network attained optimal performance, achieving 79.84% accuracy and a macro F1-score of 77.76%. It surpassed more complex architectures while maintaining the lowest parameter count of 100,324, suggesting that model parsimony is beneficial for fraud detection tasks. The multi-class framework demonstrated notable performance variability among fraud categories, with Ghost Enrollee detection attaining an outstanding F1-score of 97.84%, whereas Phantom Billing exhibited considerable classification difficulties, reflected in an F1-score of 61.15%. The

confidence analysis revealed that 57.77% of predictions attain very high confidence levels (¿0.9) with an accuracy of 97.20%, facilitating practical implementation in confidence-based automated decision-making.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The dataset used in this study is publicly available through Kaggle: "NHIS healthcare claims and fraud dataset" by Chosen, B. (2024). Available at: https://www.kaggle.com/datasets/bonifacechosen/nhis-healthcare-claims-and-fraud-dataset

Author Contributions

All authors have read and agreed to the published version of the manuscript.

- **Gaber Sallam Salem Abdalla** Methodology, validation, formal analysis, writing—review and editing, supervision.
- **Mohamed F. Abouelenein** Conceptualization, methodology, formal analysis, writing—original draft preparation, writing—review and editing, supervision, project administration.
- **Hatem M. Noaman** Software, validation, formal analysis, investigation, data curation, writing—review and editing, visualization.

REFERENCES

1. Varmedja, D., Karanovic, M., Sladojevic, S., Arsenovic, M., & Anderla, A. (2019, March). Credit card fraud detection-machine learning methods. In *2019 18th International Symposium Infoteh-Jahorina (Infoteh)* (pp. 1-5). IEEE.
2. Hancock, J. T., Bauder, R. A., Wang, H., & Khoshgoftaar, T. M. (2023). Explainable machine learning models for medicare fraud detection. *Journal of Big Data, 10*(1), 154.
3. Kittoe, J. D., & Asiedu-Addo, S. K. (2017). Exploring fraud and abuse in National Health Insurance Scheme (NHIS) using data mining technique as a statistical model. *African Journal of Educational Studies in Mathematics and Sciences, 13*, 13-31.
4. du Preez, A., Bhattacharya, S., Beling, P., & Bowen, E. (2024). Fraud detection in healthcare claims using machine learning: A systematic review. *Artificial Intelligence in Medicine*, 103061.
5. Nabrawi, E., & Alanazi, A. (2023). Fraud detection in healthcare insurance claims using machine learning. *Risks, 11*(9), 160.
6. Severino, M. K., & Peng, Y. (2021). Machine learning algorithms for fraud prediction in property insurance: Empirical evidence using real-world microdata. *Machine Learning with Applications, 5*, 100074.
7. Prova, N. N. I. (2024). Healthcare fraud detection using machine learning. In *2024 Second International Conference on Intelligent Cyber Physical Systems and Internet of Things (ICoICI)*. IEEE.
8. Bounab, R., Zarour, K., Guelib, B., & Khlifa, N. (2024). Enhancing medicare fraud detection through machine learning: Addressing class imbalance with SMOTE-ENN. *IEEE Access, 12*, 54382-54396.
9. Sumalatha, M. R., & Prabha, M. (2019, December). Mediclaim fraud detection and management using predictive analytics. In *2019 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE)* (pp. 517-522). IEEE.
10. Gupta, R. Y., et al. (2021). A Comparative Study of Using Various Machine Learning and Deep Learning Based Fraud Detection Models For Universal Health Coverage Schemes. *International Journal of Engineering Trends and Technology (IJETT), 69*(3), 96-102.
11. Wang, Z., Chen, X., Wu, Y., Jiang, L., Lin, S., & Qiu, G. (2025). A robust and interpretable ensemble machine learning model for predicting healthcare insurance fraud. *Scientific Reports, 15*(1), 218.

12. Johnson, J. M., & Khoshgoftaar, T. M. (2019). Medicare fraud detection using neural networks. *Journal of Big Data, 6*(1), 63.
13. Matloob, I., Khan, S., Rukaiya, R., Alfrahi, H., & Ali Khan, J. (2025). Healthcare fraud detection using adaptive learning and deep learning techniques. *Evolving Systems, 16*(2), 72.
14. Shah, H., Pandya, D., Panchal, K., & More, N. P. (2022, November). Classification of machine and deep learning techniques for financial fraud detection of healthcare industry. In *2022 International Conference on Futuristic Technologies (INCOFT)* (pp. 1-7). IEEE.
15. Shungube, P. S., Bokaba, T., Ndayizigamiye, P., Mhlongo, S., & Dogo, E. (2024). A Deep Learning Approach for Healthcare Insurance Fraud Detection.
16. Anand Kumar, P., & Sountharrajan, S. (2025). Insurance claims estimation and fraud detection with optimized deep learning techniques. *Scientific Reports, 15*(1), 27296.
17. Suesserman, M., Gorny, S., Lasaga, D., Helms, J., Olson, D., Bowen, E., & Bhattacharya, S. (2023). Procedure code overutilization detection from healthcare claims using unsupervised deep learning methods. *BMC Medical Informatics and Decision Making, 23*(1), 196.
18. Zhou, J., Wang, X., Wang, J., Ye, H., Wang, H., Zhou, Z., ... & Chen, W. (2023). FraudAuditor: A visual analytics approach for collusive fraud in health insurance. *IEEE Transactions on Visualization and Computer Graphics, 29*(6), 2849-2861.
19. Yoo, Y., Shin, J., & Kyeong, S. (2023). Medicare fraud detection using graph analysis: A comparative study of machine learning and graph neural networks. *IEEE Access, 11*, 88278-88294.
20. Branting, L. K., Reeder, F., Gold, J., & Champney, T. (2016, August). Graph analytics for healthcare fraud risk estimation. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (pp. 845-851). IEEE.
21. Johnson, J. M., & Khoshgoftaar, T. M. (2021). Medical provider embeddings for healthcare fraud detection. *SN Computer Science, 2*(4), 276.
22. Nugraha, R. A., Pardede, H. F., & Subekti, A. (2022). Oversampling based on generative adversarial networks to overcome imbalance data in predicting fraud insurance claim: 10.48129/kjs. splml. 19119. *Kuwait Journal of Science*.
23. Sadiq, S., & Shyu, M. L. (2019). Cascaded propensity matched fraud miner: Detecting anomalies in medicare big data. *Journal of Innovative Technology, 1*(1), 51-61.
24. Johnson, J. M., & Khoshgoftaar, T. M. (2023). Data-centric ai for healthcare fraud detection. *SN Computer Science, 4*(4), 389.
25. Hancock, J. T., & Khoshgoftaar, T. M. (2021). Gradient boosted decision tree algorithms for medicare fraud detection. *SN Computer Science, 2*(4), 268.
26. Zhang, C., Xiao, X., & Wu, C. (2020). Medical fraud and abuse detection system based on machine learning. *International Journal of Environmental Research and Public Health, 17*(19), 7265.
27. Alhassan, R. K., Nketiah-Amponsah, E., & Arhinful, D. K. (2016). A review of the National Health Insurance Scheme in Ghana: what are the sustainability threats and prospects?. *PloS one, 11*(11), e0165151.
28. Sun, J., Wang, Y., Zhang, Y., Li, L., Li, H., Liu, T., & Zhang, L. (2024). Research on the risk governance of fraudulent reimbursement of patient consultation fees. *Frontiers in Public Health, 12*, 1339177.
29. Mailloux, A. T., Cummings, S. W., & Mugdh, M. (2010). A decision support tool for identifying abuse of controlled substances by ForwardHealth Medicaid members. *Journal of Hospital Marketing & Public Relations, 20*(1), 34-55.
30. Kapadiya, K., Patel, U., Gupta, R., Alshehri, M. D., Tanwar, S., Sharma, G., & Bokoro, P. N. (2022). Blockchain and AI-empowered healthcare insurance fraud detection: an analysis, architecture, and future prospects. *IEEE Access, 10*, 79606-79627.
31. Jillo, G. (2024). Advances and Challenges in Fraud Detection in Medical Insurance. Available at SSRN 4907327.
32. Herland, M., Bauder, R. A., & Khoshgoftaar, T. M. (2019). The effects of class rarity on the evaluation of supervised healthcare fraud detection models. *Journal of Big Data, 6*, 1-33.