**International Journal of Emerging Trends in Computer Science and Information Technology**
ISSN: 3050-9246 | https://doi.org/10.63282/3050-9246.IJETCSIT-V4I1P105
Eureka Vision Publication | Volume 4, Issue 1, 34-43, 2023

*Original Article*

# Predictive Machine Learning Models for Financial Fraud Detection Leveraging Big Data Analysis

Jayakeshav Reddy Bhumireddy[1], Rajiv Chalasani[2], Srikanth Reddy Vangala[3], Ram Mohan Polam[4], Bhavana Kamarthapu[5], Mitra Penmetsa[6]
[1]University of Houston.
[2]Sacred Heart University.
[3]University of Bridgeport.
[4]University of Illinois at Springfield.
[5]Fairleigh Dickinson University.
[6]University of Illinois at Springfield.

*Abstract - Maintaining the integrity of financial systems and preventing people and organizations from suffering financial losses depends heavily on the ability to spot fraudulent financial transactions. Recognizing complicated patterns and developing fraud methods is a challenge for conventional rule-based fraud detection methods. Using the Credit Card Fraud Detection (CCFD) dataset, this research aims to compare and analyze different prediction models to accurately identify fraudulent transactions. The anonymized dataset was preprocessed thoroughly, i.e., missing values, outlier elimination, min-max normalization, and relevant feature selection were addressed. A modified Deep Neural Network (DNN) model, Multi-Layer Perceptron (MLP), Naive Bayes (NB), as well as Decision Tree (DT), were among the classification models that were trained and evaluated. The suggested DNN model performed better than the others, achieving 99.89% accuracy, a 99.87% F1-score, 99.99% recall, as well as 99% precision. These results demonstrate that deep learning, when combined with an efficient preprocessing pipeline, may greatly improve fraud detection in extremely unbalanced financial data, and they also show that the DNN model is capable of learning complicated, non-linear patterns in the data.*

*Keywords - Big Data Analytics, Financial Fraud, Fraud Detection, Machine Learning (ML), Predictive Analytics, Fraud Prevention, Transaction Data.*

## 1. Introduction

The modern finance environment has presented us with a highly engaging involvement with a fast-paced technological world and extreme dependence on online transactions. Though it has presented better convenience and speed that is unrivalled, this digital transformation has led to presenting new vulnerabilities in financial ecosystems. Financial fraud has become one of the silent but formidable enemies in this complicated network of virtual interactions. Fraudsters have often been termed as a phantom presence because they work in the shadows- they are ever changing their strategy, seeking out digital loopholes and finding ways around traditional security structures [1]. Consequently, financial fraud has evolved and become more advanced, difficult to track, and more challenging to corral with conventional rule-based detection systems.Attackers are now using sophisticated systems and programs like bots, phishing kits, and even AI-generated synthetic identities to attack transactional integrity. Financial organizations are under attack on many levels- unlawful access, data breach, identity theft, money laundering, and tampering with digital records. Financial transactions are too complex and their number is too huge to allow manual inspection or rules-based, static detection to keep up. Traditional methods of detecting fraud, which depend on strict patterns and predetermined criteria, are no longer adequate[2][3]. The need of the hour is a paradigm shift toward adaptive and intelligent systems that can understand transaction behaviours in real time and detect novel fraudulent patterns as they emerge. ML and AI have emerged as a solution to these evolving threats for researchers and industry professionals.

The technologies introduce the force of automation, pattern recognition, and continual learning. There are several forms of monetary fraud, including fraud involving credit cards, financial statements, bankruptcy, insurance, telecom subscriptions, and stocks. The application of ML-based techniques to address these categories has been the subject of several studies, with models SVM, Decision Trees, Neural Networks, and Ensemble Learning being used. Thus, prior research has used the features of transactions (time, location, type of merchant, spending pattern of a user, etc.) to train classification models that are effective at identifying anomalies [4][5][6]. Moreover, behavioural analytics, clustering, and hybrid models have been applied by the researchers to improve the fraud detection systems and make them compatible with the evolving fraud tactics [7].One of the most important achievements in this field is the combination of predictive ML models and big data analytics. Such systems do not just classify frauds, but also detect the ones that may happen in the future by finding hidden patterns in large and varied data. With big data, these models can handle millions of transactions in real time, and can mix structured and unstructured data sources, including user logs, geolocation, transaction histories and device metadata [8][9]. Such a combination enables more subtle and situation-specific decision-making. With the use of such capabilities, contemporary fraud detection systems are getting more proactive, scalable, and resilient. The paper discusses these predictive ML methods used in financial fraud detection as one of the key areas where big data analysis is helping policymakers and stakeholders to manoeuvre and combat the continuously changing world of financial crime.

### *1.1. Motivation and Contribution*

The exponential growth of online financial transactions has made credit card fraud a serious problem that poses serious hazards to both consumers as well as financial institutions. Conventional fraud detection techniques frequently have high false-positive rates and can't keep up with changing fraud trends. This study is motivated by the need for an intelligent, accurate, and scalable solution that can proactively detect fraudulent behavior in real-time. Leveraging deep learning (DL) through DNNs provides an opportunity to build a more resilient fraud detection system that can learn from complex data patterns with minimal manual feature engineering. This study makes a substantial contribution to network safety in several ways:

- Developed a DL-based CCFD system with a structured pipeline of preparation that incorporates data normalization, outlier removal, and handling of missing values.
- Successfully addressed class imbalance using balancing techniques, significantly enhancing model performance and fraud recall.
- Achieved outstanding evaluation metrics accuracy, proving the accuracy and consistency of the model for detecting fraud in real-time.
- The suggested DNN model outperformed the conventional models (DT, NB, as well as MLP) in terms of robustness and predictive capacity, according to a thorough comparison study.

### *1.2. Justification and Novelty*

The novelty of this work is, the study incorporates a DNN-based architecture with extensive data preprocessing strategies to boost the identification of credit card fraud. Unlike traditional models, the proposed approach incorporates robust steps like class balancing, outlier removal, and feature selection to improve model training and generalization. The exceptional performance marked accuracy and nearly perfect recall, justifies the use of DNNs, as they are capable of capturing intricate, non-linear patterns inherent in financial transaction data. This methodology addresses common pitfalls like data imbalance and overfitting, making it a significant advancement in fraud detection research.

## 2. Literature Review

Several significant research studies on financial fraud detection have been thoroughly reviewed and analyzed to guide and strengthen the development of this study, some of the related studies are discussed and summarized in table 1 below.Gardner et al. (2019) focus on detecting financial fraud by developing a three-tiered anomaly detection system. More than one random forest classifier with varying fitness functions is tuned to build the system. The process optimizes the random forest parameters to satisfy the fitness function via a randomized grid search. Afterwards, the models are contrasted to provide three levels of detected fraud, where each level represents a different level of accuracy. By isolating identified frauds into various levels, it is feasible to get both great recall as well as precision. The technique achieves an accuracy rate of 96% when classifying frauds and a high precision of over 90% when detecting 85% of frauds. Tiered random forest outperforms other algorithms, according to studies, with a recall of 72% and a precision of 85%, when compared to SVM and logistic regression [10].

Adepoju et al. (2019) identified several factors that greatly impact the effectiveness of CCFD, including knowledge of online payment fraud, data set measurement approach, variables used, and detection methodologies. Using highly skewed credit card fraud data, this publication applies SVM, LR, NB, and KNN. Specificity, accuracy, sensitivity, and precision are the four main metrics used to assess the efficacy of these approaches. LR achieved an optimal accuracy of 99.07%, KNN 96.91%, NB 95.98%, as well as SVM classifiers 97.53%. LR performs better than competing methods, according to the relative results [11].Mubalaike and Adali, (2018) focus on learning how DL models may help identify fraudulent transactions with more accuracy. The data was retrieved from a month's worth of authentic financial records kept by an African mobile money provider. These logs comprised more than six million transactions. Data that has previously been preprocessed is utilized with the top DL methods, which include stacked auto-encoders, and ML approaches, like ensemble decision trees, to create classifiers. The specificity, accuracy, confusion matrix, and precision. Additionally, the performance of the built classifier models is assessed using their ROC values. The optimum accuracy outcomes are 91.53%, 80.52%, and 90.49%, respectively. The results of the comparison show that the constrained Boltzmann machine outperforms the other methods[12].

Shiguihara-Juarez and Murrugarra-Llerena (2018) aim to be significant for society and financial institutions. To identify fraud, supervised ML approaches were used. On these issues, however, mostly discriminatory methods were used. Fraud may also be detected using probabilistic graphical models, which likewise show the reasoning process as a graph. They suggested a way to use domain-related constraints to create a probabilistic graphical model for fraud detection. They surpassed previous baseline methods of probabilistic graphical models and obtained an accuracy of 99.272%. They showed that to address difficult issues like fraud detection, restrictions are crucial [13]. Chouiekh and El Haj (2018) used DL algorithms as a successful way to identify mobile communications scammers. Extracting learning characteristics was made possible by using fraudulent datasets obtained from the customer data records of a legitimate wireless communication service and classifying event behavior as either fraudulent or non-fraudulent. The proposed model was evaluated using a battery of tests. Deep convolution neural networks (DCNN) achieved an impressive 82% accuracy rate, much outperforming more conventional ML algorithms such as RF, SVM, as well as Gradient Boosting Classifier, as well as training time. Consequently, this approach can reduce costs related to unauthorized service usage [14].Awoyemi, Adetunmbi and Oluwadare (2017) analyze the effectiveness of KNN, LR, and NB on highly skewed CCFD. There are 284,807 transactions in the dataset, which is derived from credit card holders throughout Europe. The biased data is processed using a mixed method that incorporates both under- and over-sampling techniques. The unprocessed and processed data sets undergo the three procedures.

The system's backbone is Python. To measure how well the approaches work, look at their accuracy, sensitivity, precision, specificity, rate of classification balance, and Matthews correlation coefficient. The data shows that LR classifiers reach an optimal accuracy rate of 54.86%, KNN obtains 97.69%, and NB achieves 97.92%. In terms of performance, k-nearest neighbours is superior to both NB and LR methods [15].Lin et al. (2015) collect data using data mining algorithms and an expert questionnaire to identify potential fraud indicators and follow up by ranking the various factors based on their importance. Classification and Regression Trees (CART), & Artificial Neural Networks (ANNs) are some of the data mining techniques being used in the study. The ANNs and CART methods achieve a 91.2 percent accuracy rate in classifying training and test data (ANNs) and 90.4 percent (CART), and 92.8 percent (ANNs) and 90.3 percent (CART) respectively much precise compared to the logistic model that has only 83.7 percent and 88.5 percent of the correct classification in predicting the occurrence of the fraud. Besides, ANNs type II error minimizes to 23.9 percent, 27.8 percent, and 43.3 percent compared to CART and logistic models [16].

Table I provides an overview of current research on financial fraud detection, focusing on the state-of-the-art models that have been utilized, datasets used, major findings, and the challenges faced in each case.

**Table 1. Overview Of Recent Studies On Machine Learning Models For Financial Fraud Detection**

| Author | Proposed Work | Dataset | Key Findings | Challenges/recommendations |
|---|---|---|---|---|
| Gardner et al. 2019 | Three-tiered anomaly detection using tuned Random Forest classifiers with different fitness functions | Custom dataset | 96% frauds correctly classified; >90% precision for 85% of detected frauds | Tiered random forest achieves a better balance of precision and recall than SVM or logistic regression |
| Adepoju et al. 2019 | Tested KNN, NB, LR, and SVM on skewed fraud datasets | CCFD Dataset (Kaggle / distorted version) | Logistic Regression achieved 99.07% accuracy; others between 95–97% | Performance relies heavily on variable choice and dataset skew handling |
| Mubalaike and Adali, 2018 | Used ensemble decision trees, stacked auto-encoders, and RBM on mobile money data | Mobile Money Transaction Logs (6 M+ transactions from an African company) | RBM achieved the highest accuracy (91.53%) | DL is effective on real-time, large-scale financial logs |
| Shiguihara-Juarez & Murrugarra-Llerena, 2018 | Developed a probabilistic graphical model with domain-specific constraints | Simulated fraud detection dataset | Achieved 99.272% accuracy, better than baseline PGMs | Domain constraints improve fraud detection performance |
| Chouiekh and El Haj, 2018 | Used DNNs for telecom fraud detection on a large scale | Customer Detail Records (CDR) from a real mobile carrier | DCNN achieved 82% accuracy, outperforming traditional models | DCNN offers better detection accuracy and reduced cost for telecom fraud |

| | | | | |
|---|---|---|---|---|
| Awoyemi, Adetunmbi & Oluwadare, 2017 | Use hybrid sampling to evaluate NB, KNN, as well as LR on data that is skewed | European Credit Card Fraud Dataset (Kaggle - 284,807 transactions) | LR: 54.86%, KNN: 97.69%, NB: 97.92%; KNN was the most effective. | Hybrid sampling (over + under) improves the model on skewed datasets |
| Lin et al., 2015 | Combined expert questionnaires with LR, CART, and ANN for fraud detection | Real banking dataset | ANN showed the best classification (92.8%) and reduced Type II errors | ANNs are more accurate than traditional LR; significantly lower false negatives |

## 3. Research Methodology

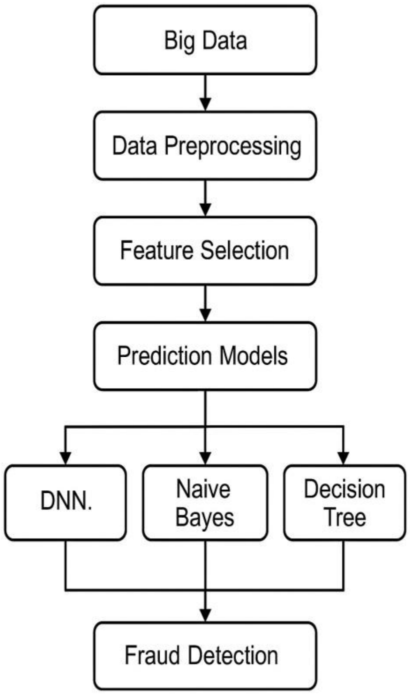

**Figure 1. Proposed Flowchart for Financial Fraud Detection**

The methods presented in this study are based on a DNN model to systematically tackle the problem of CCFD. It starts with the pre-processing of the dataset, including the missing values imputation, and outliers removal, and continues with the class imbalance addressment by balancing methods. Data normalization, feature selection, and data partitioning into training and test sets follow. Fitting the DNN model to the training data allows us to evaluate the test data using accuracy, F1-score, precision, as well as recall. The end product of such a pipeline is a working fraud-detecting system. With this systematic procedure, the model will perform better because data quality will be enhanced, and the skewed classes will be handled. The model allows representing the intricate patterns in the transaction behaviour by using DL, which can be applicable in real-time fraud prediction. The overall workflow emphasizes data-driven refinement to increase robustness and reliability in financial fraud detection. Figure 1 shows the entire procedure. Each step in the suggested flowchart for creating ML models targeted at financial fraud detection is thoroughly defined as follows.

### *3.1. Data Collection*

The CCFD Dataset is part of the collection. The data comes as a result of the Kaggle ML platform. It is a dataset that contains 3075 transactions in 12 features of transactions in CSV format. The feature details and the background information cannot be presented due to confidentiality concerns. The images of data visualization (bar plots, heatmaps, etc.) helping to explore the distribution of attacks, correlations between features, etc., are presented below:

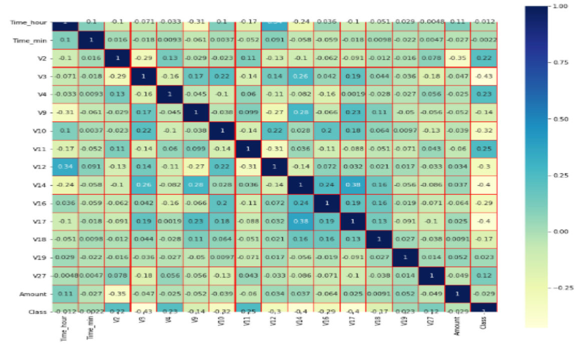

**Figure 2. Correlation Heatmap of Credit Card Fraud Detection**

Figure 2. The heatmap displays the correlation matrix of various features in a dataset, including time-related features (Time_hour, Time_min), anonymized variables (V1 to V28), Amount, as well as the target variable (Class). Each feature is completely associated with itself. therefore, all the diagonal values are 1. Most off-diagonal correlations are low (close to 0), indicating minimal linear relationships between different features. Notably, the target variable "Class" shows very weak correlations with all other features, which is typical in fraud detection datasets and suggests that predicting the target class is inherently difficult. The matrix also highlights some mild to moderate correlations among features like V2, V4, V10, and V14. The requirement for more sophisticated methods, such as non-linear models, to successfully uncover concealed patterns is highlighted by this sparse correlation pattern.

### *3.2. Data Pre-Processing*

The CCFD Dataset was gathered, cleaned, and concatenated, and pertinent characteristics were extracted as part of the data preparation process. Pre-processing was done on the dataset to eliminate missing values and Exceptions. Pre-processing, the data transformation and normalization, is carried out. The following steps of pre-processing are as follows:

- **Handling missing value:** Deletion involves removing the incomplete rows, while imputation involves replacing the missing data with a model or another statistical measure of central tendency, such as the mean or median. It is essential to guarantee high-quality data for training ML models.
- **Remove Outliers:** ML algorithms employ the outlier removal approach to exclude data points that substantially differ from the rest of the dataset since they may have an impact on a classifier's performance.

### *3.3. Handling Class Imbalance*

The dataset showed a clear disparity between transactions that were fraudulent and those that were not. The Synthetic Minority Over-Sampling Technique (SMOTE) was applied to the training data after the train-test split, and a solution was generated. A key component of SMOTE's strategy for minority representation is the creation of fictional minority class members. This prevents data from leaking into the test set while the model learns minority class patterns.

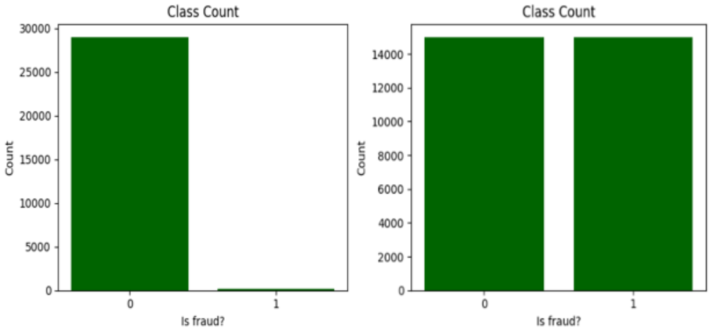

**Figure 3. Class Distribution Before and After Balancing for Fraud Detection Dataset.**

Figure shows two bar charts illustrating class distribution for the "Is fraud?" label before and after data balancing. The left Figure shows a huge class imbalance, with the bulk of non-fraudulent instances (label 0) far outnumbering fraudulent ones (label 1). In contrast, the right chart depicts a balanced dataset where both classes have nearly equal representation, demonstrating the effective use of a data balancing method to correct the class imbalance, such as oversampling or SMOTE.

### 3.4. Data Normalization

The data was normalized using the min–max technique, which restricted the values to a range of 0 to 1. This was done to maximize the performance of the classifiers involved and to reduce the impact of outliers. The normalization followed the below mathematical Equation (1): $X'$ $X_{min}$ $X_{max}$

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

The original value of the feature is represented by X, the normalized value by $X'$, the minimum value by $X_{min}$, and the maximum value by $X_{max}$.

### 3.5. Feature Selection

The accuracy of a model may be improved by feature selection, decreasing overfitting, and enhancing interpretability by eliminating duplicate or unnecessary features. The purpose of feature selection is to identify the most informative qualities of a dataset for use in the construction and training of an ML model. Feature selection improves the performance of the AI model by dimensionally restricting the feature space to a selected subset, hence decreasing its processing needs.

### 3.6. Data Splitting

The dataset was divided into training and testing sections to assess the efficacy of the model in issue; 80 percent of the data was to be used for parameter estimation and model training, while the remaining 20 percent was designated for testing and calculating model performance.

### 3.7. Proposed Deep Neural Network (DNN) Model

DNN models are a type of ANN that can learn and represent complicated data structures due to the numerous hidden layers that exist between the input as well as output layers. Every layer has a set of connected neurons that perform a weighted computation and an activation function that is not linear, like sigmoid or ReLU. Optimization techniques that minimize a loss function, like cross-entropy or mean squared error, are learnt by DNNs. Examples of such algorithms include Adam and SGD. Such models have been applied extensively in image recognition, natural language processing, and speech analysis applications because they have a high feature extraction and pattern recognition capacity. DNN represents an improved form of the traditional ANN, having a minimum of 3 layers that are hidden. To fully understand how DNN works, a good grasp of the principles of artificial neural network is needed. The following formula (2) determines the DNN output:

$$y(t) = \sum_{k=1}^{L} f\big(w_k + x_k(t)\big) + \in (t)$$

In which $w_k$ are the weights of the layer trained by backpropagation $x_k$ (k = 1, ..., L) is the number of sequences of real values known as events, within an epoch. f is the activation function.

### 3.8. Evaluation Metrics

The effectiveness of the proposed architecture was examined using several performance metrics. The actual values and the anticipated results of trained models were contrasted. False-Negatives (FN), False-Positives (FP), True-Negatives (TN), and True-Positives (TP) were evaluated using this comparison. Recall, accuracy, precision, F1-score, and confusion are all included in the following matrix. The following explains it:

**Accuracy:** A measure of how well the trained model predicted outcomes relative to the whole dataset (input samples). It is given as Equation (3)-

$$Accuracy = \frac{\mathrm{TP+TN}}{\mathrm{TP+Fp+TN+FN}}$$

**Precision:** Precision is measured as the proportion of accurate results to all correct and incorrect results. How good the classifier is in predicting the positive classes is expressed as Equation (4)-

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP+FP}}$$

**Recall:** The percentage of accurate predictions compared to the total number of right and incorrect ones is known as the recall statistic. In mathematical form it is given as Equation (5)-

$$Recall = \frac{\mathrm{TP}}{TP+FN}$$

**F1 score:** In other words, it aids in maintaining a healthy equilibrium between recall and accuracy by combining the two. Its range is [0, 1]. It is mathematically expressed as Equation (6)-

## 4. Results and Discussion

The experimental setup and the suggested model's performance during training and testing are provided in this section. The Python programming language was used to run this model on a computer running Microsoft Windows 10 with a 2.30 GHz processor and 8GB of RAM. The outcomes of using the CCFD Dataset to train the suggested model are displayed in Table II. and testing it with a crucial performance matrix that includes recall, precision, accuracy, as well as F1-score. Using the CCFD dataset, the proposed DNN model for financial fraud detection achieved remarkable results across all evaluation metrics. The model attained an extremely high accuracy of 99.89%, which shows its general efficiency in making the right predictions. The DNN accuracy of 99% indicates its high capability of reducing the false positive rate, whereas the recall of 99.99% indicates its almost perfect capacity in predicting the real fraud instances. Moreover, an F1-score of 99.87 percent indicates a balanced model between recall as well as precision, and it proves that the developed model is robust and reliable in identifying fraudulent activities.

**Table 2. Experiment Results of Proposed Models for of Financial Fraud Detection on Credit Card Fraud Detection dataset**

| Performance Matrix | Deep Neural Network (DNN) Model |
|---|---|
| Accuracy | 99.89 |
| Precision | 99 |
| Recall | 99.99 |
| F1-score | 99.87 |

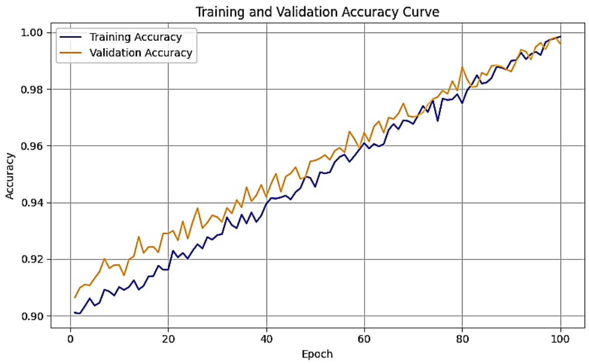

**Figure 4. Accuracy curves for the DNN Model**

Figure 4 displays, over the course of 100 iterations, the accuracy curve of the proposed DNN model for validation & training. The upward trend in both the curves is steady and therefore reflects successful learning and good generalization ability. Although smaller fluctuations are noticed, as is normal in the training process, the general performance continues to improve, achieving an accuracy rate of more than 99.9% throughout training and validation. The near parallelism of the two curves ascertains minimum overfitting and is an indicator of the strength and stability of the model when carrying out tasks related to detecting credit card fraud.

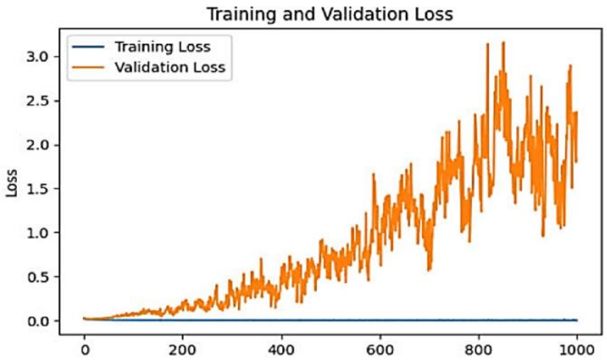

**Figure 5. Loss Curves for the DNN Model**

Figure 5 demonstrates the loss (validation and training) on 1000 epochs. The model appears to be fitting the training data because the training loss (blue line) is flat and extremely low. Conversely, the validation loss (orange line) grows tremendously with time, becoming extremely erratic and peaking at above 3.0 at the end. A model that learns the training data but is unable to generalise to new data is said to be severely overfitting, as seen by the growing difference between the training and validation losses. This indicates poor model generalization and a need for regularization techniques or early stopping.

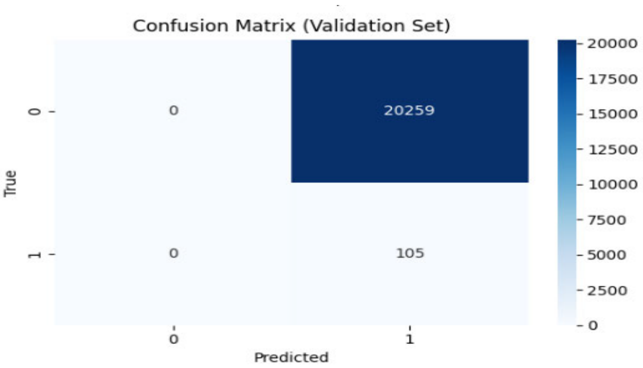

**Figure 6. Confusion Mcatrix for DNN Model**

Figure 6 shows that the matrix of misunderstanding for the validation set shows a severe class imbalance and poor predictive performance. Class 0 was incorrectly predicted as class 1 20,259 times, while class 1 was accurately forecasted as class 1 105 times; the model also forecasts all samples as class 1. There are zero true negatives and false negatives, indicating the model never predicts class "0" at all. Possible causes for this bias include an imbalance in the training data or insufficient model training, since the model appears to be biased towards class "1". Despite the model's apparent great accuracy, its applicability is seriously questioned.

### *4.1. Comparative Analysis*

A comparison accuracy evaluation is conducted with various existing models to guarantee the suggested DNN model produces superior results. Table III compares the accuracy of several prediction models that were employed to identify financial fraud on the CCFD dataset as part of this investigation. The DT model demonstrated an accuracy value of 72%, which indicates a relatively low performance measured against other models. The Naïve Bayes (NB) classifier showed much better results and an accuracy level of 95.99 %. The MLP measured more accurately than NB, at 96.4%. The DNN surpassed all others with an accuracy value of 99.89 %, illustrating its advanced suitability to detect financial fraud, in comparison to the other models.

**Table 3. Accuracy Comparison of different Predictive models for Financial Fraud Detection using the CCFD Dataset**

| Models | Accuracy |
|---|---|
| DT[17] | 72% |
| NB[11] | 95.99 |
| MLP[18] | 96.4 |
| DNN | 99.89% |

The proposed DNN model has many advantages in financial fraud detection. The model has the benefit of automatic training, allowing it to learn complex, non-linear patterns from large datasets, thereby being able to detect subtle and more sophisticated fraudulent actions that other models may not have been able to discover. Strong predictive power is demonstrated with the DNN model achieved 99.89% for real-time classification on the CCFD data set. The robustness of the DNN is potentially backed up as well, as the model could provide predictive power on any dataset once trained on another dataset in the financial domain. The auto-extraction of features through the DNN deep architecture contributes to the lack of manual feature extraction and provides for scalability across datasets of varied financial data. Consequently, the DNN model has become a reliable and vigilant tool for fraud detection.

## 5. Conclusion and Future Study

Robust risk management has emerged as a key initiative in the expansion and development of the financial sector. Hence, the use of ML-based approaches to develop risk control models is gaining currency in more and more financial institutions. This paper According to the results of several predictive models applied to the CCFD dataset, it can be concluded that data pre-processing essentially influences the work of the model. Pre-processing procedures, such as missing value removal, outlier removal, min-max normalization, and feature selection procedures, made the dataset clean, scaled, and finally dataset was optimized for training. The steps played an important role in making the data ready to be effectively learned, considering the nature of the dataset itself, including the imbalance between classes and poor feature relationships. DNN was found to be the best model of all the tested models, with an accuracy of 99.89, beating DT, NB, and MLP classifiers. It proves that with the help of strong pre-processing, state-of-the-art models such as DNN can be very effective in detecting fraudulent transactions in highly imbalanced financial data.Future research will investigate the inclusion of other ML and DL models, including ensemble methods and transformer-based architectures, to improve detection accuracy and models' robustness against undesirable maintenance effects. A hybrid model that includes traditional algorithms and The trade-offs between predictive capability & processing efficiency will also be examined using deep networks. Future research will equally examine real-time detection frameworks, explainable AI schemes that promote user understanding of model decisions, and its integration with sufficient validation with larger and heterogeneous datasets in financial institutions to enhance scalability and accountability in applied financial modelling.

## References

[1] S. Zhang et al., “HiDDen: Hierarchical Dense Subgraph Detection with Application to Financial Fraud Detection,” in Proceedings of the 2017 SIAM International Conference on Data Mining, Philadelphia, PA: Society for Industrial and Applied Mathematics, 2017, pp. 570–578. doi: 10.1137/1.9781611974973.64.

[2] N. B. Harikrishnan, R. Vinayakumar, K. P. Soman, and Prabaharan Poornachandran, “Time split based pre-processing with a data-driven approach for malicious URL detection,” in Advanced Sciences and Technologies for Security Applications, 2019. doi: 10.1007/978-3-030-16837-7_4.

[3] V. Kolluri, “A Thorough Examination of Fortifying Cyber Defenses: AI in Real Time Driving Cyber Defence Strategies Today,” Int. J. Emerg. Technol. Adv. Appl., 2018.

[4] D. Zhang and L. Zhou, “Discovering golden nuggets: Data mining in financial application,” IEEE Trans. Syst. Man Cybern. Part C Appl. Rev., 2004, doi: 10.1109/TSMCC.2004.829279.

[5] A. Singh and A. Jain, “An empirical study of AML approach for credit card fraud detection-financial transactions,” Int. J. Comput. Commun. Control, 2019, doi: 10.15837/ijccc.2019.6.3498.

[6] V. Kolluri, “An In-Depth Exploration of Unveiling Vulnerabilities: Exploring Risks in AI Models and Algorithms,” Int. J. Res. Anal. Rev., vol. 1, no. 3, pp. 910–913, 2014.

[7] S. Wang, “A comprehensive survey of data mining-based accounting-fraud detection research,” in 2010 International Conference on Intelligent Computation Technology and Automation, ICICTA 2010, 2010. doi: 10.1109/ICICTA.2010.831.

[8] J. West and M. Bhattacharya, “Intelligent financial fraud detection: A comprehensive review,” 2016. doi: 10.1016/j.cose.2015.09.005.

[9] S. Garg, “Predictive Analytics and Auto Remediation using Artificial Inteligence and Machine learning in Cloud Computing Operations,” Int. J. Innov. Res. Eng. Multidiscip. Phys. Sci., vol. 7, no. 2, 2019.

[10] C. Gardner, D. C.-T. Lo, J.-C. Chern, P. Paschos, and C. Ng, “Tiered Financial Fraud Detection Utilizing Precision Stratified Random Forest Assembly,” in 2019 IEEE 5th International Conference on Big Data Intelligence and Computing (DATACOM), 2019, pp. 254–257. doi: 10.1109/DataCom.2019.00047.

[11] O. Adepoju, J. Wosowei, S. Lawte, and H. Jaiman, “Comparative Evaluation of Credit Card Fraud Detection Using Machine Learning Techniques,” in 2019 Global Conference for Advancement in Technology, GCAT 2019, 2019. doi: 10.1109/GCAT47503.2019.8978372.

[12] A. M. Mubalaike and E. Adali, “Deep Learning Approach for Intelligent Financial Fraud Detection System,” in UBMK 2018 - 3rd International Conference on Computer Science and Engineering, 2018. doi: 10.1109/UBMK.2018.8566574.

[13] P. Shiguihara-Juarez and N. Murrugarra-Llerena, “A Bayesian Classifier Based on Constraints of Ordering of Variables for Fraud Detection,” in 2018 Congreso Internacional de Innovacion y Tendencias en Ingenieria, CONIITI 2018 - Proceedings, 2018. doi: 10.1109/CONIITI.2018.8587081.

[14] A. Chouiekh and E. H. I. EL Haj, “ConvNets for Fraud Detection analysis,” Procedia Comput. Sci., vol. 127, pp. 133–138, 2018, doi: 10.1016/j.procs.2018.01.107.

[15] J. O. Awoyemi, A. O. Adetunmbi, and S. A. Oluwadare, “Credit card fraud detection using machine learning techniques: A comparative analysis,” in Proceedings of the IEEE International Conference on Computing, Networking and Informatics, ICCNI 2017, 2017. doi: 10.1109/ICCNI.2017.8123782.

[16] C. C. Lin, A. A. Chiu, S. Y. Huang, and D. C. Yen, “Detecting the financial statement fraud: The analysis of the differences between data mining techniques and experts’ judgments,” Knowledge-Based Syst., 2015, doi: 10.1016/j.knosys.2015.08.011.

[17] S. Patil, V. Nemade, and P. K. Soni, “Predictive Modelling for Credit Card Fraud Detection Using Data Analytics,” in Procedia Computer Science, 2018. doi: 10.1016/j.procs.2018.05.199.

[18] A. Thompson, L. Aborisade, O. Oyinloye, and E. Odeniyi, “A Fraud Detection Framework using Machine Learning Approach,” Fourth Int. Conf. Cyber Technol. Cyber Syst. 2019., pp. 12–18, 2019.

[19] Polu, A. R., Buddula, D. V. K. R., Narra, B., Gupta, A., Vattikonda, N., & Patchipulusu, H. (2021). Evolution of AI in Software Development and Cybersecurity: Unifying Automation, Innovation, and Protection in the Digital Age. Available at SSRN 5266517.

[20] Chinta, P. C. R., Katnapally, N., Ja, K., Bodepudi, V., Babu, S., & Boppana, M. S. (2022). Exploring the role of neural networks in big data-driven ERP systems for proactive cybersecurity management. Kurdish Studies.

[21] Routhu, K., Bodepudi, V., Jha, K. M., & Chinta, P. C. R. (2020). A Deep Learning Architectures for Enhancing Cyber Security Protocols in Big Data Integrated ERP Systems. Available at SSRN 5102662.

[22] Chinta, P. C. R., & Katnapally, N. (2021). Neural Network-Based Risk Assessment for Cybersecurity in Big Data-Oriented ERP Infrastructures. Neural Network-Based Risk Assessment for Cybersecurity in Big Data-Oriented ERP Infrastructures.

[23] Katnapally, N., Chinta, P. C. R., Routhu, K. K., Velaga, V., Bodepudi, V., & Karaka, L. M. (2021). Leveraging Big Data Analytics and Machine Learning Techniques for Sentiment Analysis of Amazon Product Reviews in Business Insights. American Journal of Computing and Engineering, 4(2), 35-51.

[24] Kalla, D. (2022). AI-Powered Driver Behavior Analysis and Accident Prevention Systems for Advanced Driver Assistance. International Journal of Scientific Research and Modern Technology (IJSRMT) Volume, 1.

[25] Chinta, P. C. R. (2022). Enhancing Supply Chain Efficiency and Performance Through ERP Optimisation Strategies. Journal of Artificial Intelligence & Cloud Computing, 1(4), 10-47363.

[26] Kuraku, D. S., Kalla, D., & Samaah, F. (2022). Navigating the link between internet user attitudes and cybersecurity awareness in the era of phishing challenges. International Advanced Research Journal in Science, Engineering and Technology, 9(12).

[27] Sadaram, G., Sakuru, M., Karaka, L. M., Reddy, M. S., Bodepudi, V., Boppana, S. B., & Maka, S. R. (2022). Internet of Things (IoT) Cybersecurity Enhancement through Artificial Intelligence: A Study on Intrusion Detection Systems. Universal Library of Engineering Technology, (2022).

[28] Karaka, L. M. (2021). Optimising Product Enhancements Strategic Approaches to Managing Complexity. Available at SSRN 5147875.

[29] Polu, A. R., Vattikonda, N., Buddula, D. V. K. R., Narra, B., Patchipulusu, H., & Gupta, A. (2021). Integrating AI-Based Sentiment Analysis With Social Media Data For Enhanced Marketing Insights. Available at SSRN 5266555.

[30] Jha, K. M., Bodepudi, V., Boppana, S. B., Katnapally, N., Maka, S. R., & Sakuru, M. Deep Learning-Enabled Big Data Analytics for Cybersecurity Threat Detection in ERP Ecosystems.

[31] Kalla, D., Smith, N., Samaah, F., & Polimetla, K. (2022). Enhancing Early Diagnosis: Machine Learning Applications in Diabetes Prediction. Journal of Artificial Intelligence & Cloud Computing. SRC/JAICC-205. DOI: doi. org/10.47363/JAICC/2022 (1), 191, 2-7.

[32] Kalla, D., Kuraku, D. S., & Samaah, F. (2021). Enhancing cyber security by predicting malwares using supervised machine learning models. International Journal of Computing and Artificial Intelligence, 2(2), 55-62.

[33] Katari, A., & Kalla, D. (2021). Cost Optimization in Cloud-Based Financial Data Lakes: Techniques and Case Studies. ESP Journal of Engineering & Technology Advancements (ESP-JETA), 1(1), 150-157.

[34] Kalla, D., Smith, N., Samaah, F., & Polimetla, K. (2021). Facial Emotion and Sentiment Detection Using Convolutional Neural Network. Indian Journal of Artificial Intelligence Research (INDJAIR), 1(1), 1-13.