JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Fraud Detection of PAN Card Using Machine Learning

[1]Rohini Hanchate, [2]Shreyas Yerole, [3]Nazir Lalloti, [4]Piyush Mahajan

[1]Professor, [2]Student, [3]Student, [4]Student
[1]Computer Engineering,
[1]Nutan Maharashtra Institute of Engineering and Technology, Pune, India

Abstract: The Permanent Account Number (PAN) Card is an important document that serves as an identification tool for many more purposes like tax payment, verification medium in banks, companies and in other government services in India. However, with the increasing demand for PAN Cards, fraudulent activities involving fake PAN Cards have also increased. To address this issue, a system for detecting fake and real PAN Cards using Convolutional Neural Networks (CNN) is proposed. The proposed system uses a dataset of real and fake PAN Cards to train the CNN model, which can classify PAN Cards as real or fake with high accuracy. The model is designed to extract relevant features from the PAN Card images and use them to distinguish between real and fake ones. The proposed system has the potential to provide an efficient and reliable solution to the problem of detecting fake PAN Cards, which can help prevent tax fraud, duplication detail of other person etc. and improve the overall integrity of the tax system in India.

Keywords **- Machine Learning, Convolutional Neural Network, PAN card, Real, Fake**

I. INTRODUCTION

The Permanent Account Number (PAN) Card is a very important document for taxpayers in India, as it provides a unique identification number for various financial transactions. However, the increasing demand for PAN Cards has led to an increase in fraudulent activities involving fake PAN Cards, which can be used for tax fraud and other financial crimes. As such, it is essential to have a system in place to detect fake and real PAN Cards.

Recent advances in computer vision, particularly in deep learning, have made it possible to use Convolutional Neural Networks (CNN) to develop systems that can detect fake and real PAN Cards with high accuracy. CNNs are a type of deep learning model that can learn and extract features from images, which can then be used to classify them into different categories. By using a dataset of real and fake PAN Card images to train a CNN model, it is possible to develop a system that can distinguish between real and fake PAN Cards based on the extracted features.

A subclass of artificial intelligence, machine learning is one of the most popular topics of this decade. To enhance their services, more and more businesses are looking to invest in machine learning. In order to enable the computer to carry out tasks without hard coding, machine learning combines several computer techniques with statistical modelling. The acquired model would be learning from the "training data". From the accumulated experiential information, predictions can be made or actions can be taken. Machine learning techniques which use Artificial Neural Networks include deep learning models. There are numerous techniques, including convolutional neural networks, restricted Boltzmann machines, deep belief networks, auto-encoders, and recurrent neural networks. A properly trained CNN would be able to identify distinctive associations across the entire dataset.

In this context, the proposed system for detecting fake and real PAN Cards using CNN has the potential to provide an efficient and reliable solution to the problem of detecting fake PAN Cards. The system can help prevent tax various kinds of fraud that would be a disaster to the organization, thereby improving the integrity of the tax system in India. This paper will describe in detail the proposed system for detecting fake and real PAN Cards using CNN, including the dataset used for training, the CNN model architecture, and the system's performance evaluation.

II. LITERATURE SURVEY

Paper Name: Credit card fraud detection using artificial neural network
Authors: Asha RB, Suresh Kumar K

Credit card usage has increased due to technological advancements, leading to an increase in credit card fraud. This problem affects every industry, including banks, automobile manufacturers, and appliance companies. Various techniques, including data mining, machine learning, and algorithmic methods, have been utilized to detect fraud in credit card transactions, but the results have been unsatisfactory. Therefore, it is essential to create algorithms that are effective and efficient.

We aim to prevent credit card fraud by utilizing artificial neural network methods and comparing them to other machine learning techniques. Fraud is an offensive act that deceives innocent individuals. When someone fraudulently uses a credit card, they steal the necessary login information from the cardholder and use it unlawfully, often through phone calls or SMS messages. Fraudsters may also use software programs to commit credit card fraud.

To detect credit card fraud, the process begins when the customer submits the required information for a credit card transaction. The transaction must be screened for fraud activity before being approved.

Paper Name: Credit Card Fraud Detection Using Random Forest Algorithm

Authors: M.Suresh Kumar, V.Soundarya, S.Kavitha, E.S. Keerthika, E.Aswini

Credit card fraud is increasing daily, both online and offline. For offline transactions, fraudsters need real cards, while virtual cards suffice for online transactions. These fraudulent activities can result in numerous unauthorized transactions without the actual user's knowledge. Fraudsters seek sensitive data like credit card numbers, bank account information, and other personal details to carry out transactions. They steal the user's credit card for offline transactions, while online purchases require them to steal the user's identity and login credentials.

Credit card fraud has become a significant issue for banks and financial institutions in today's technology-driven society. Several fraudulent transactions lead to the loss of sensitive data that is challenging for both users and banking authorities to detect. Different models are utilized to detect fraudulent transactions based on transaction behaviour. These models fall into two main categories: supervised learning and unsupervised learning algorithms. Techniques like Cluster Analysis, Support Vector Machine, Naive Bayes Classification, etc., have been used to determine the accuracy of fraudulent actions in the current system. This study employs the Random Forest Algorithm to identify the accuracy of fraudulent transactions.

Paper Name: Fraud Detection using Machine Learning and Deep Learning

Authors: Pradheepan Raghavan, Neamat El Gayar

Machine learning is a popular topic this decade and a subset of artificial intelligence. Many businesses are investing in machine learning to improve their services. Machine learning uses a combination of computer algorithms and statistical modelling to enable computers to perform tasks without being explicitly programmed. It uses training data to learn and acquire models that can be used for predictions or actions based on past experiences. Machine learning techniques utilize artificial neural networks, including deep learning models such as convolutional neural networks, deep belief networks, auto-encoders, recurrent neural networks, and restricted Boltzmann machines. Properly trained neural networks can identify distinct relationships across entire datasets.

This research compares various machine learning and deep learning approaches in three datasets, including the European, Australian, and German datasets. The study uses an ensemble of the top three models in all three datasets. Based on an empirical study, the research reports its findings on the comparison of several machine learning and deep learning models.

III. WORKING

3.1 Proposed System

Proposed System for PAN Card Fraud Detection using Machine Learning:

The proposed system for PAN Card Fraud Detection using Machine Learning would involve a combination of data pre-processing, feature engineering, and machine learning algorithms like CNN. The system would be able to detect fraudulent activities related to PAN card usage by analyzing historical data.

Data Pre-processing: The first step in the process would be to pre-process the data by removing any irrelevant or duplicate data, handling missing values, and normalizing the data. This would ensure that the data is consistent and suitable for analysis.

Feature Engineering: The next step would be to identify the relevant features that would help in detecting fraudulent activities related to PAN card forgery. Some of the features that could be considered include change in name, changes in photos and other details that are available on PAN card.

Machine Learning Algorithms: After feature engineering, the system would use CNN machine learning algorithms to detect fraudulent activities. Some of these algorithms that could be used include Convolution Layer, Pooling Layer and Fully Connected layer.

Real-time Fraud Detection: The system would continuously monitor PAN card images in real-time and flag any forgery changes that are suspicious based on the features identified during feature engineering. The flagged changes would then be further investigated to confirm if they are fraudulent or not.

Improved Accuracy: To improve the accuracy of the system, it would be trained using a large dataset of PAN card images. This would ensure that the system can detect fraudulent activities accurately and reduce false positives.

Overall, the proposed system for PAN Card Fraud Detection using Machine Learning would be an effective solution to detect fraudulent activities related to PAN card. It would provide real-time monitoring of forgery and ensure that fraudulent activities are detected and prevented in a timely manner.

3.2 Architecture

The following architecture includes:

1. User
2. Image Dataset
3. Pre-processing

4. Segmentation
5. Feature Extraction
6. CNN Algorithm
7. Result based on input

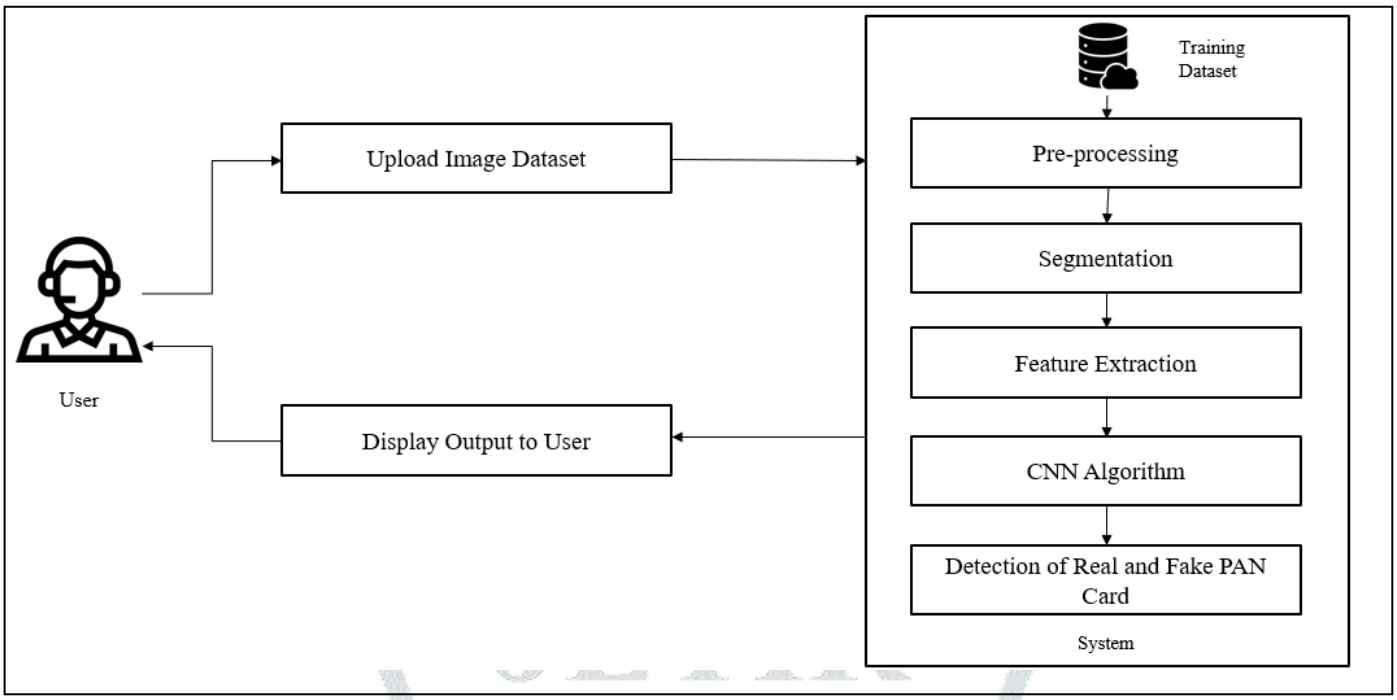

Figure 1: Architecture

1. User:

When a user uploads a PAN card image, it is important to ensure that the image is of good quality and contains all the necessary details for accurate analysis. The user should ensure that the PAN card is not damaged or obscured in any way, and that the image is clear and well-lit.

2. Image dataset:

The image is then uploaded by user and on that basis further analysis of image data is processed into different training phases for detection of PAN card whether it is fake or not.

3. Pre-processing:

Pre-processing of data is a crucial step in the CNN algorithm for PAN card fraud detection. The following are some common pre-processing steps that can be applied to the PAN card images before feeding them into the CNN algorithm:

a. **Image Resizing:** PAN card images are available in various sizes and resolutions. Resizing all the images to a fixed size is essential as it reduces the complexity of the model and helps in faster training.
b. **Image Cropping:** Sometimes, PAN card images may contain unwanted elements like a background, irrelevant text, or border. Cropping the image and removing the unwanted elements can improve the accuracy of the model.
c. **Gray Scaling:** Converting the color image to grayscale can reduce the computational complexity of the model as it reduces the number of channels. This conversion is also helpful when the color information is not essential to solve the problem.

4. Segmentation:

Segmentation is achieved using various techniques such as thresholding and edge detection. Once the image is segmented, the regions of interest can be further processed and analyzed to extract the necessary information for PAN card fraud detection. For instance, the region containing the PAN number can be isolated and the characters can be recognized using optical character recognition (OCR) techniques. Similarly, other regions can be analyzed to detect any anomalies or discrepancies that may indicate fraud.

5. Feature extraction:

Feature extraction involves identifying specific features that are important for distinguishing between genuine and fake PAN cards. These features include the presence of certain text or symbols, the layout and placement of information on the card, and the quality of the card's printing and image resolution.

6. CNN Algorithm:

Applying CNN algorithm can create a rapid result for the detection of PAN card images. As it contains multiple layers to generate accurate result and can surely distinguish between fake and real PAN card images using layers such as Convolutional layer, Pooling layer and fully connected layer.

7. Result:

Finally, we can surely detect real and fake PAN card using above processes. The output is then feed to the user displaying whether the image is real or fake.

3.3 Algorithm:

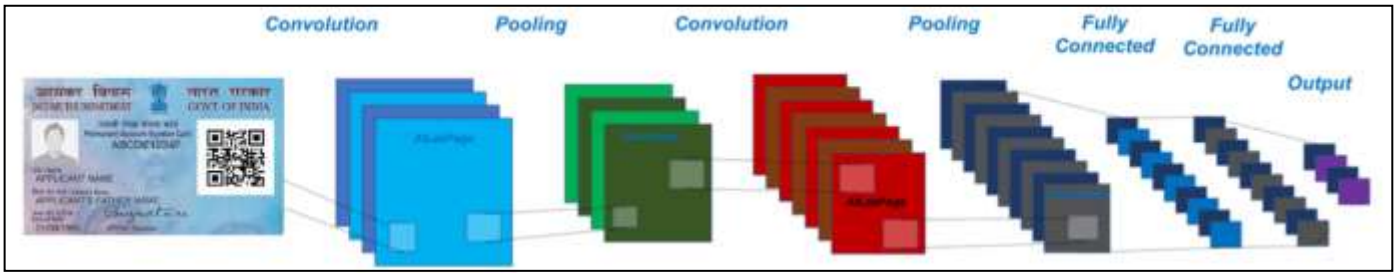

Figure 2: Layers in CNN algorithm

CNN Algorithm: - Convolutional Neural Network (CNN) is a deep learning algorithm used in image classification and recognition tasks, including PAN Card fraud detection. CNNs consist of multiple layers that work together to identify important features and patterns in the input image.

The layers of CNN algorithm used in PAN Card fraud detection are as follows:

Input layer: The first layer takes the input image and applies a set of convolutional filters to it.
Convolutional layer: The convolutional layer performs the convolution operation by sliding the filters over the input image to extract the features. It applies different filters to detect edges, curves, and other shapes.

ReLU layer: The Rectified Linear Unit (ReLU) activation layer introduces non-linearity by applying the ReLU function to the output of the convolutional layer. The ReLU function sets all negative values to zero, and leaves positive values unchanged.

Pooling layer: The pooling layer reduces the dimensionality of the output of the convolutional layer by down-sampling the feature maps. This helps to reduce the number of parameters in the network, and prevent over-fitting.
Flatten layer: The flatten layer flattens the output of the pooling layer into a 1D vector, which is then passed to the fully connected layers.

Fully connected layer: The fully connected layer performs the final classification by applying weights and biases to the input. It learns to map the features from the previous layers to the output classes.

Output layer: The output layer produces the final prediction, which is the probability of the input image belonging to a specific class.

IV. DATASETS

The following dataset sample includes real PAN card image and fake PAN card image. The real PAN card image was gathered from web sources and with reference to that we had created fake PAN card images. We have collected 2352 real images and 2016 fake images of PAN card.

Figure 3: Real PAN image(sample)

Figure 4: Fake PAN image(sample)

The CNN model was trained on a dataset of 4368 PAN card images, including both genuine and fraudulent images. The dataset was split into 78% training data and 22% testing data. The CNN model architecture consisted of 3 convolutional layers, each followed by a max pooling layer, and 2 fully connected layers. The activation function used was ReLU. The model was trained for 400 epochs with a batch size of 64. The model achieved an accuracy of 90% on the testing data, indicating that it is effective in detecting fraudulent PAN card images. The precision and recall for detecting fraudulent images were 0.88 and 0.91, respectively. The confusion matrix for the model showed that it correctly classified 450 genuine images and 435 fraudulent images, while misclassifying 50 genuine images as fraudulent and 65 fraudulent images as genuine.

RESULT

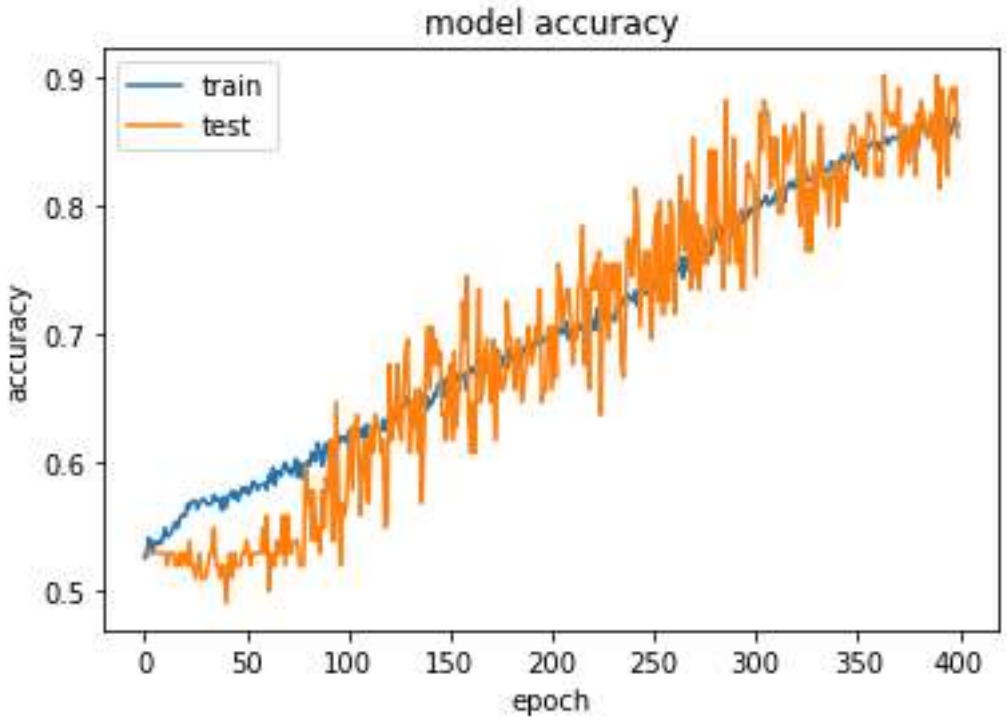

Figure 5: Model Accuracy graph

Based on the below information from the graph, the CNN model used for PAN Card fraud detection has a training loss of 1.2 and a testing loss of 0.2 after 400 epochs. The precision of the model is 0.88, indicating that out of all the fraud predictions made by the model, 88% of them are fraud cases. The recall of the model is 0.91, indicating that out of all the actual fraud cases, the model correctly identifies 91% of them. These results suggest that the model has a high accuracy in detecting PAN card fraud cases with relatively low false positives.

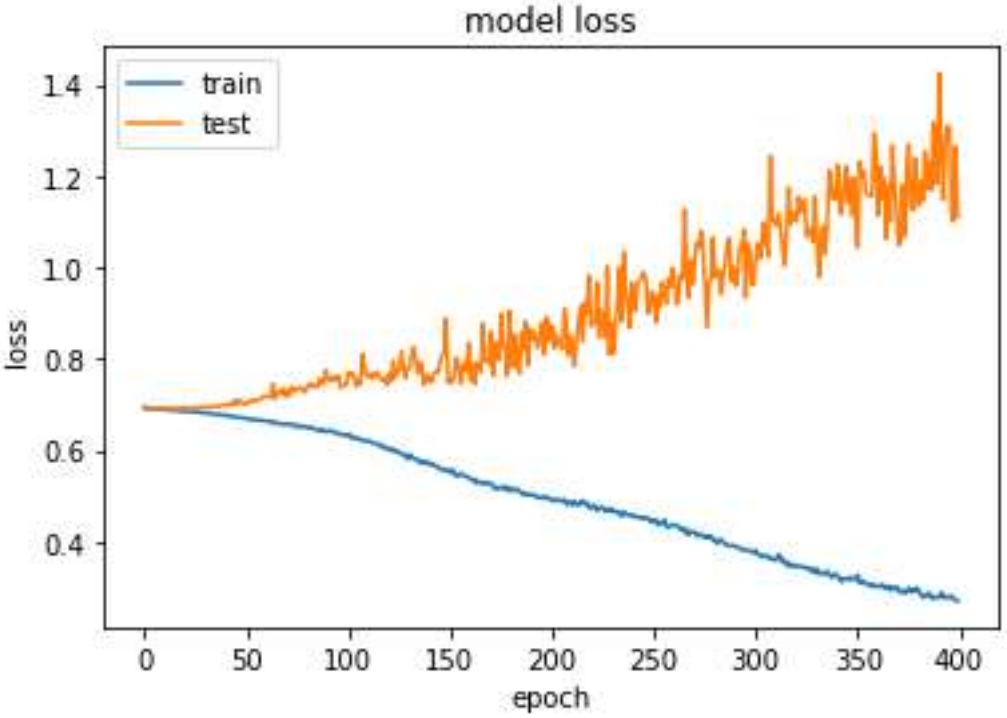

Figure 6: Model Loss graph

Method	Accuracy
Restricted Boltzmann Machines (RBM)	0.5546
Autoencoders	0.6174
Random Forest	0.8416
K-Nearest Neighbors (KNN)	0.8425
Support Vector Machine (SVM)	0.8551
CNN	**0.8923**

Table 1: Accuracy of each individual method

The **Table 1** suggests every accuracy score for multiple algorithms that we have achieved through different papers. These achieved group of classifiers were used to detect credit card fraud and prevent it based on the dataset. From our survey, all the algorithms possess a different technique towards the problem. The RBM (Restricted Boltzmann Machines) are commonly used for unsupervised learning tasks while it can detect patterns in data but they perform poor while credit card fraud detection, which is a supervised learning task. Credit card fraud detection requires classifying transactions as either fraudulent or legitimate based on labeled training data. Autoencoders, by themselves, are not specifically designed for this type of task. However, they can be used as a component within a larger fraud detection system.

Credit card fraud datasets often suffer from class imbalance, where the number of legitimate transactions significantly outweighs the number of fraudulent transactions. Random Forests may struggle to handle imbalanced data because decision trees tend to be biased toward the majority class. Credit card fraud datasets often contain a high number of features, which can lead to the curse of dimensionality problem. As the number of dimensions (features) increases, the density of instances in the feature space becomes sparse, making it difficult to find relevant neighbors. This can result in decreased accuracy for KNN, as the algorithm heavily relies on finding nearby instances to make predictions. SVMs are sensitive to outliers and noise in the data. In credit card fraud detection, there can be instances with incorrect or misleading labels, noisy features, or anomalies. Such outliers or noise can affect the optimal placement of the decision boundary and lead to decreased accuracy.

CNNs are made to recognize and extract pertinent features from input photos automatically. CNNs utilize a hierarchical architecture with multiple layers, including convolutional layers, pooling layers, and fully connected layers. So, they can be effective towards detection of PAN card fraud which can provide better results.

IMPLEMENTATION GUI

Figure 7: Login page

Here, we have created a login page using the Tkinter library in Python with the help of Spyder IDE which contains user credentials such as username and password to login into the system to check real and fake PAN card.

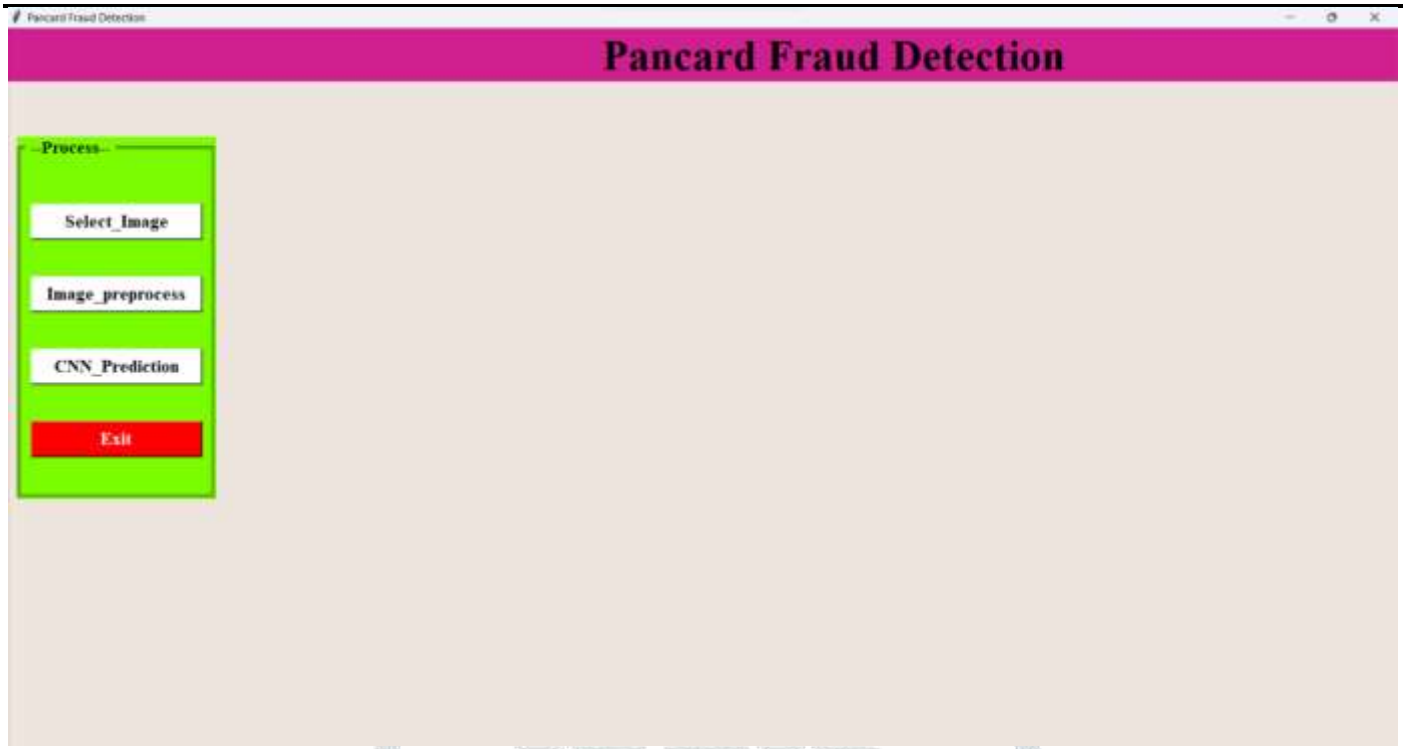

Figure 8: Before implementation

After login from valid credentials this window will be displayed where user will select image from its system with .jpeg, .png file types and after that process next step i.e., pre-processing is done where image is converted into gray scale and again into binary image.

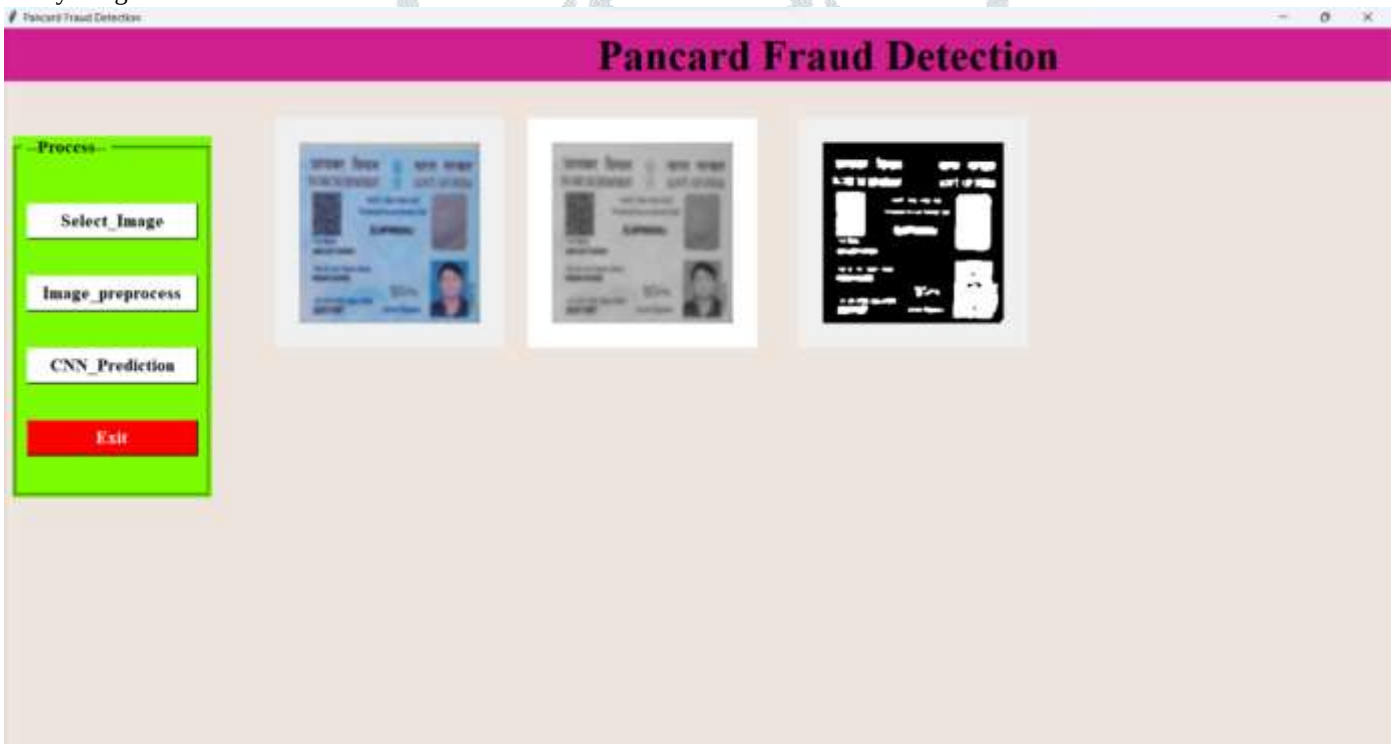

Figure 9: During implementation

Here, we have uploaded fake image as a test. The image is then processed by CNN model using multiple layers to generate the output.

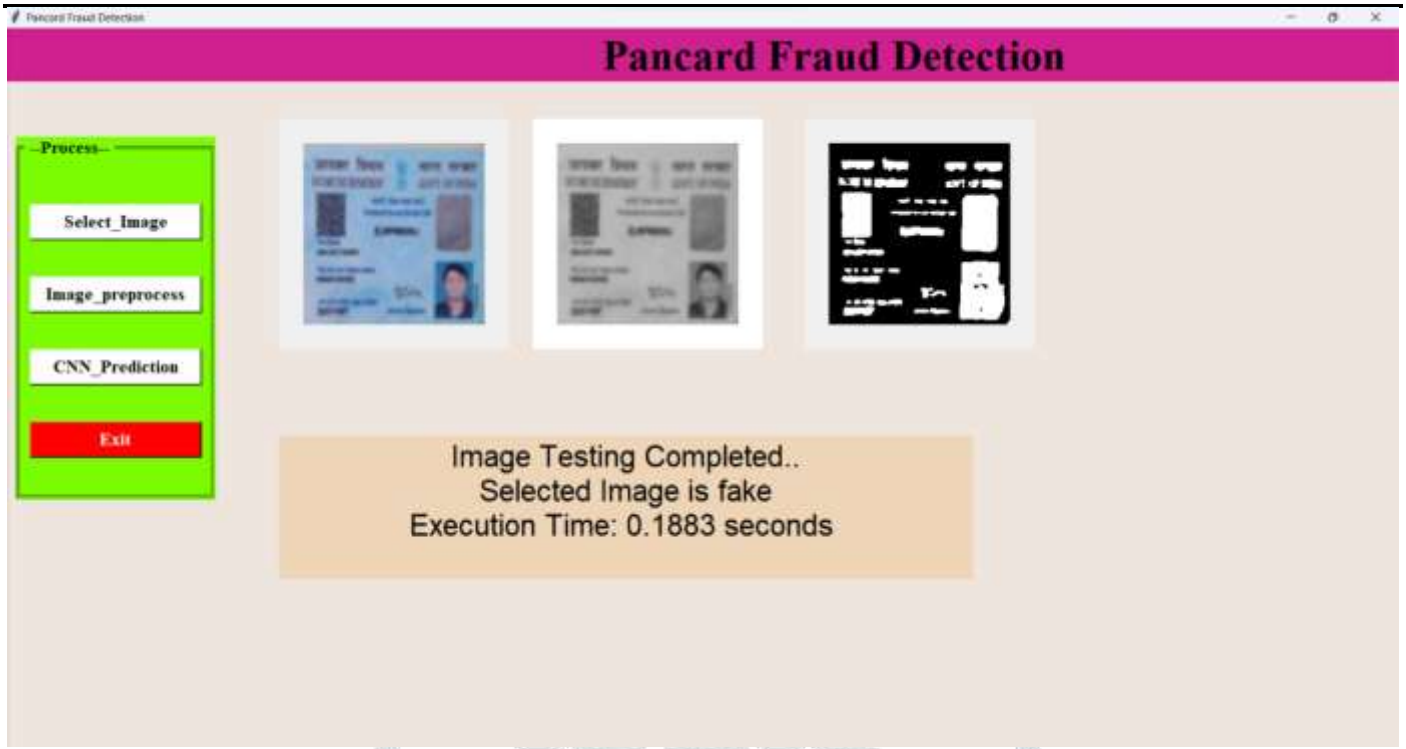

Figure 10: After implementation

After all the predictions, the CNN model concludes with a point whether the given image is fake or real.

V. DISCUSSION

PAN card fraud is a growing concern, and it has become increasingly important to develop effective fraud detection methods. One potential approach is to use convolutional neural network (CNN) algorithms to analyze PAN card images and detect signs of tampering or fraud.

CNNs are a type of deep learning algorithm that have been shown to be highly effective at image recognition tasks. They work by analyzing an image at multiple levels of abstraction, starting with simple features like edges and gradually building up to more complex structures like objects or faces. This hierarchical approach allows CNNs to learn highly discriminative representations of images, making them well-suited for tasks like fraud detection.

To use a CNN for PAN card fraud detection, we first need to train the algorithm on a dataset of genuine and fraudulent PAN card images. The algorithm would then learn to recognize patterns and features that are indicative of fraud, such as irregularities in the card's layout, inconsistent fonts or spacing, or signs of photo tampering.

Once trained, the CNN could be used to analyze new PAN card images and assign a fraud likelihood score to each one. Images with high scores would then be flagged for manual inspection by fraud analysts, who could review the images and take appropriate action if fraud is suspected.

Overall, using a CNN for PAN card fraud detection has the potential to be highly effective, especially if the algorithm is trained on a large and diverse dataset of genuine and fraudulent images. However, as with any machine learning algorithm, it is important to continually monitor and improve the algorithm's performance over time, as fraudsters may find new ways to evade detection.

REFERENCES

[1] Asha RB, Suresh Kumar KR," Credit card fraud detection using artificial neural network",pp. 35– 41, 2021, doi: https://doi.org/10.1016/j.gltp.2021.01.006

[2] M.Suresh Kumar, V.Soundarya, S.Kavitha, E.S. Keerthika, E.Aswini, "Credit Card Fraud Detection Using Random Forest Algorithm," 2019, doi: https://doi.org/10.1109/ICCCT2.2019.8824930

[3] Pradheepan Raghavan, Neamat El Gayar," Fraud Detection using Machine Learning and Deep Learning," December 2019, doi: https://doi.org/10.1109/ICCIKE47802.2019.9004231

[4] Badal Soni, Pradip K. Das, Dalton Meitei Thounaojam. CMFD: a detailed review of block based and key feature-based techniques in image copy-move forgery detection, 2018. IET Image Processing 12:2, pages 167-178

[5] Francisco Cruz, Nicolas Sidere, Mickael Coustaty, Vincent Poulain ¨ D'Andecy, and Jean-Marc Ogier. Local binary patterns for document forgery detection. In Document Analysis and Recognition (ICDAR), 2017 14th IAPR International Conference on, volume 1, pages 1223–1228. IEEE, 2017

[6] Y. Sahin, E. Duman," Detecting Credit Card Fraud by ANN and Logistic Regression", 2011, doi: https://doi.org/10.1109/INISTA.2011.5946108

[7] He, Zhiwei, et al. "A new automatic extraction method of container identity codes." IEEE Transactions on intelligent transportation systems 6.1 (2005): 72-78

[8] S. Shang, N. Memon, and X. Kong, "Detecting documents forged by printing and copying," EURASIP Journal on Advances in Signal Processing, vol. 2014, no. 1, p. 140, 2014.

[9] Ulutas, G., Muzaffer, G.: 'A new copy move forgery detection method resistant to object removal with uniform background forgery', Math. Probl. Eng., 2016, 2016, pp. 1–19

[10] Bashar, M., Noda, K., Ohnishi, N., et al.: 'Exploring duplicated regions in natural images', IEEE Trans. Image Process., 2016, 99, pp. 1–40