CREDIT CARD FRAUD DETECTION

1.
INTRODUCTION
Nowdays, online payment methods have been used widely as effect of the quicks
increase in non-cash electronic transaction. Credits cards are one of the electronic
payments method A credit card is a thins rectangular shapepiece of plastics or metals
released by a bank or financial services company to a customers (cardholder) to
facilitate payment to a supplier of goods and service. It is based on the consumers.The
cards issuer (usually a bank) open an accounts, which is generally circlings, and
contributes a
line of credit to the users. Which the users can use to make a payments. With a cardbased payments reporting for approximately 51% of transactions. [1], [2], [3]. Despite
the
advantages of electronic payments, credit card companies are experiences an increase
in card fraud with the beginning of many new technologies. Scammers are smart
sufficient to takes advantage of excuses and always try to steal data using new
technologies like Skimming and phishing. There are occurrence when a website is
designed to match a legitimate sites and victims enter personal information such as

password, user name, and credit card informations The fraudster send out a major
number of emails that direct victims to their bogus websites. The e-mails seems to be
from company such as AOL, PayPal banks and eBay, and they ask the victims to log
their personal informations in order to resolve issues." The fraudsters earnings by theft
the victim's identities and then theft their money [4]. Credit card fraud caused a heavy
financial loss from card frauds. According to a 2017 US Payments Forum reports,
criminals have shifted their
focus to movements involving CNP transactions as
chip cards security has improved‖ [5]. The estimated financials loss of credit cards
fraud credit cards fraud worldwide in 2018 rose to $24.26 million‖[6] 2018 rose to
$24.26 million‖ [6]. By 2019, the worldwide frauds losse had accounted for US $27
billion accordings to PR Newswire Association LLC‖.[7] Moreover, it is estimated that it
will surpass around $30 billion by 2020‖ [8]. Activation procedures have all contributed
to a reduction in the effect of
fraud. Merchants are putting programs in place to help
prevent credit card fraud. although, more safety measures must be taken to prevent frauds [4]. Fraudulent transactions are well detected with the help of
(ml)Machine Learning algorithms that have a high processing or computing powers and
the ability to handle large datasets. which is a promising way to reduce credit cards
frauds [9], [10]. This paper includes seven ections. Section II summarizes brief previous

studies. Sections III the approache by which the basic studie were systematically
chosen are offered. In sections IV. Several
popular credit cards fraud detection
methods have been riefed. Section V. presented a comparison of various fraud
detection methods. Section IV. Summarizes results and discussion. Finally section VII.
Presented conclusion and future scope.

1.1 Credit Card Fraud

Fraud according to the Organization of Certified are defined asany malicious or
deliberate acts of depriving another of ownerships or money through wiliness, cheating,
or other unfair means [11].The unauthorized procedure of CC or information destitute of
owners data is called add the full name and then the abbreviations CCF. The different
CCF trick applications & behaviors are related to two groups of frauds. Specify the first
group and the second group. When app frauds occurs, fraudsters apply for new cards
from the bank or provide it to companies that use false or other information. A customer
can file multiple applications with a single usuals of describes (named duplicate fraud),
or a different customers with similar describes (named identity fraud). Instead, there are
practically four
main types of behavioural frauds: stolen/lost cards, mail thefts,fake cards, & current
cardholder does not exist‘ frauds. When a stolen / lost card fraud occurs, fraudsters

steals a credit cards or get lost card. Mail theft frauds when a fraudster receives
personal details from a bank in the mail before a credit cards or original card holder.
Fake & Card Holders Frauds & credit cards descriptions are not presented. In past,
remote communications can be done using card details via mail,phone or internet.
Second,(where is first) fake cards are created on cards data" explain more here [12].

1.2 Credit Card Fraud Detection

Service make electronic payments more easy, seamless, adequate, and simple to use;
however, we must not overlooks the losse associated with electronic commerce.
company and banks to use them offer good security solutions. To address these issues,
but fraudster’s delicate techniques evolves over time. As a result, it is criticals toimproving detections and preventions method’s [7]. It is critical to understand the
mechanisms for carrying a frauds in order to fight the fraud effectively. The gadget for
identifying credit score card fraud depend upon on the fraud manner itself [13].
To accomplish this, provides the transactions details to the verification segment which
will classify them as either fraud or non-fraud. If it classified as fraudulents, it will be
refused.Otherwise, the transaction is accepted [14]. Fraud detections methods such as
statistical data analysis and artificial intelligence can be used to differentiate between

the two. AI method’s consist of data mining that used to detects frauds, which can
classify, groups and modules data to search through millions of transactions to find
patterns and detect frauds. (MI)Machine learning is a methods for automatically
detecting
fraud characteristics. One technique of dealing with frauds is through both prevention and detections. Fraud detection and prevention's primary goal
is to tell difference between legitimate and fraudulent transactions and to prevents
fraudulent action. Using historical data, the user's pattern and behaviores are analysed
to verify if a transaction is fraudulent or not. When the systems fails to detects and
prevents fraudulent activities, fraud detection takes over. [15]. In supervised fraud
detection systems, new transactions are classified as fraudulent or authentic based on
descriptions of deceptive and legitimate activities, whereas outliers' transactions are
identified as prospective fraudulent
transactions in unsupervised fraud
detection systems. A point by-point dialogue between supervised and unsupervised
machine learning techniques can be discovered. Diversity of research have been
conducted on severals methods to solve the problems of credit cards fraud detection.
These approaches include,ANN, K-means Clustering, DT, etc.[16].

1.3 Frauds type in Card-based transactions

1) Physical Cards Fraud in most POS (point of sale) transactions, as it is essentials that
the customer’s must have to be physically presenting the cards to the merchants to
carry out the transactions. There are chances that the cardholders cardscan be stolen
and misuse by fraudsters without the cardholder knowledge.
2)Virtual Card Fraud: In
most Online shopping transaction there are no need for a physical card and instead we
use the Card Numbers, CVV number ,and Expiry Date, to perform the transactions.
Fraudsters can steal this details and they can use it to perform fraudulent online
transactions‖ [17].2. LITERATURE REVIEW
Prajal Save et al. [18] have proposed a model based on a decision tree and a
commbinations of Luhn's and Hunt's
algorithms. Luhns algorithms is used to
determine whether
incoming transactions is fraudulent or not. It validates credit card numbers via the input,
which is the credit cards number. Address Mismatch and Degrees of Outlierness are
used to assess the deviation of each incoming transactions from the customer’s normal
profile. In the final step, the general belief is strengthened or weakened using Bayes
Theorems, followed by recombination’s of the calculated probability with the initial belief
of frauds using an advanced combination heuristic.Vimala Devi. J et al. [19] To detect
counterfeit transactions,three (ML)machine-learning algorithms were presented and

implemented. There are many measure used to evaluate the performances of classifier
or predictors such as the Vector Machine, Random Forest, and Decision Tree. These
metrics are either prevalence-dependent or prevalence-independent.Furthermore these
technique are used in credit card fraud detections mechanisms, and the results of these
algorithms have been compared. Popat and Chaudhary [20] supervised algorithms were
presented Deep learning, Logistic Regression, Nave Bayesian, Support Vector Machine
(SVM), Neural Network, Artificial Immune System, K Nearest Neighbour, Decision
Tree ,Data Mining, Fuzzy logic based System, and Genetic Algorithm are some of the
techniques used. Credit cards fraud detections algorithms identify transactions that
have a high probability of being fraudulent's.We compared (ml)machine-learning
algorithms to prediction, clustering, and outlier detection. Shiyang Xuan et al. [21] For
training the behavioral characteristics of credit cardtransaction the Random Forests
classifier was used. The following types are used to trains the normal and fraudulent's
behavior feature Random forest based on randoms trees and random forest based on
CART. To assess the model's effectiveness, performances measures are computed.
Dornadula and Geetha S. [5] Using the Sliding-Window method, the transactions were
aggregated into respectives groups , some feature from the window were extracted to

find customer’s behavioral patterns. Features such as the maximum amount, the
minimum amount of a transaction, the average amounts in the windows and even the
time elapsed are available. Sangeeta Mittal et al. [22] To evaluate the underlying
problems, some popular machine learning algorithms in the supervised and
unsupervised categories were selected. A range of supervised learning algorithm from
classical to latest, have been considered. These include tree-based algorithms,
classical and deep neural networks, hybrid algorithms and Bayesian approaches. The
effectiveness of (ml)machine-learning algorithms in detecting credit card fraud has been
assessed. On various metrics, a numbers of popular algorithms in the supervised,
ensemble, and unsupervised categories were evaluated. It is concluded that
unsupervised algorithms handle datasets skewness better and thus performs well
through all metrics absolutely and in comparison to other techniques. Deepa and Akila
[17] For frauds detection, different algorithms like Anomaly Detection Algorithm, KNearest Neighbor, Random Forest, K-Means and Decision Tree were used. Based on a
given scenario, presented severals techniques and predicted the best algorithmsto detect deceitful transactions. To predict the fraud results, the systems used various
instructions and algorithms to generate the Frauds score for that certain transactions.
Xiaohan Yu et al. [23] have proposed a deep network algorithms for fraud detections A

deep neural network algorithms for detecting credit cards frauds was described in the
papers. It has described the neural network algorithms approach as well as deep neural
network application. The preprocessing method's and important loss for resolving data
skew problems in the datasets. Siddhant. Bagga et al. [24] presented many techniques
for determining whether a transactions is real or fraudulent Evaluated and compared the
accomplishments of 9 techniques on data of credit cards fraud, including logistic
regression, KNN, RF, quadrant discriminative

3. RESEARCH METHODOLOGY

Systematic literature review, for example, are a type of methodology, which conducts a
literature reviews on a specific topic, could be used to detect frauds. A systematic
reviews primary goals in this context is to identify, evaluates, and Interprets the
available studies in the literature that address the Author's research questions. A
secondary goal is to identify research gap and opportunities in the area of interests. In
this papers, we attempted to walk through the activities proposed by Kitchenham:
analysis preparation, execution, and reporting in iterations. [28].

3.1 Selection of rudimentary Studies

To highlights primary research for selection, keywords werepassed to the search engine

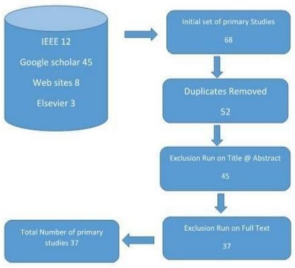
then they were chosen to enhance the development of researchs that wishes to aid in
answering the study question. The only Boolean factors thatcould be used were AND
and OR. (‖machine-learning‖ OR Artificial intelligenc) AND —fraud detection‖ were the
search terms. IEEE Explore Digitals Library was one of the platforms looked into.
-Google Scholar
- Elsevier- Science Direct
According on the search platform's the title keywords, and abstract were all searched
for. On March 28, 2021, we conducted the searche and we went over all of the previous
studies. The outcomes of these searche refined using the criteria described in Section
3.2, resulting in a collections of result that could be run.
3.2 Inclusion and Exclusion Criteria
Modern technological fraud detections, Case studies,research and comment on how to
improve existing mechanism's by creating a hybrids approach could all be considered
for inclusion in this SLR. Papers must be reads and writes in the English language. Any
Google Scholar findings and tested for submissions as if Google Scholars has the ability
to re-turn lower-grade papers. This SLR will only accepts the most recent version of a
samples.3.3 Selection Result
The primary keyword searches against the pick platforms yielded 68 studies. After
duplicate studie were removed, this was reduced to 52. After the procedures of the
survey through the implication/exception criteria, there were 45 papers left to read. The

45 papers have been read in their entirety, after applying the inclusion and exclusion
criteria a second time 37 papers remained. As a results SLR will comprise 37 papers in
total, as illustrated in the diagrams4.

CREDIT CARD FRAUD DETECTION TECHNIQUES

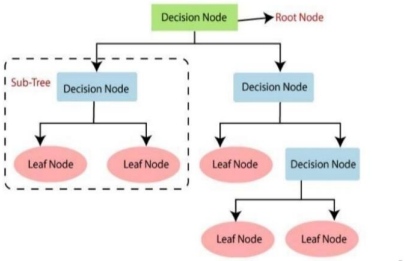
4.1 Decision tree

A supervised learning methodology, graphical representations of possible solution's to



choices based on certain situation's [29] As in Figure and it is a tree-structured lassifier.
It starts with a roots node where inside nodes represent the features of a
datasets,branches symbolize the decision instructions and each leaf node representsthe results. In decision tree and they have the purpose of deciding and communicatings
espectively. A decision tree plainly asks a questions and then divide it into
sub trees based on the answers. Although DT can solve classifications and regression
problems. it is most commonly used to solve classification problems. To find the
datasets classes, the algorithm searches at the top of the tree. It compares the root
Trait with the record attribute and follow the offshoot on way to the next node, which it

calculate depending on the relations [30].Step Working Of Decision Tree
In the first phase, starts with S, which is the root node and includes the entire datasets.
Second, discovers the best Trait in the datasets using the Attribute Selection Measure.
When the node can't be categorized, in that time the final node is called a foliate node.
Based on the labels, the root node is extra subdivided into the decisions node and one
leaf node. In the end the nodes is divided into two leaves .



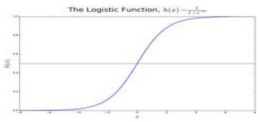
4.2 Random Forest
Random Forest classifiers finds decision trees in a subsets of the data and then
aggregates their details to that to get the full datasets predictive powers. Rather than
relying on a single decision tree. The RF take the predictions from each tree and
forecasts the final outputs based on the majority votes of forecasts. Using a huge
number of trees in the forest improves precisions and eliminates the issues of over
fitting. It predicts output with high precisions, and it runs efficiently even with large
datasets. It can also keeps accuracy when large proportions of data is lost. Random

Forest can handle both classification and regression task. It can handle largedatasets with high dimensionality. It improves the models accuracy and avoids the over
fitting problems. We use two-step training techniques in the process of tree-based
Random Forest: First, we generate the random forest by mixing N tree togethers, and
then we estimate for each of the trees we generate in the first phase [31]. An ensemble
algorithms employs the "random forest" artificial intelligence techniques. Because it
avert over fitting by averaging the results this approach outperforms single decision tree.
Random Forest is
an ensemble of diverse tree, similar to Gradient Boosted Tree, but unlike GBT, RF tree
grow in parallel. Random Forest have a lot of uncorrelated trees. Because various trees
are trained in parallel, the overall model diminishes a large numbers of variances.
Random Forest treat each trees as a separate classifiers that has been trained on
resampled data. As a results of employing this this learn strategy and divide, the
models overall learning ability are increased [10], [32].The Random Forest Working Step
These step illustrate Figure above; in the first step, choose (K) as data points at
random from the drill sets. Second, constructs the DT linked with the chosen data points
(Subsets). Following that select the digit (N) for the numbers of decision trees you wish
to construct. Then, duplicate Steps 1 and 2. Finally, discover the predictions of each

decision trees for new data point and assigns the modern data point to the category
that receives most votes. Clarify how RF work by using the following scenario: Assume
you have a datasets with a variety of fruites images. As outcome, RF classifiers will be
given this datasets. Each decision tree is given a portion of the datasets to deal with.
When a new data point occur's, the Random Forest classifier predicts the conclusion
based on the majority of outcome's

4.3 Logistic RegressionAn algorithm that can be used for both regression and classification task but it is most
commonly used for classification. Logistic Regression is used to predicts categorical
variables using dependent variables. Consider two classes, and a new data point is to
be checked to see which class it belongs to. The algorithms then compute probability
values ranging between (0) and (1). Logistic Regression employs a more complex cost
functions, this cost functions is
known as the Sigmoid Function or Logistic Function. [33]. LR also does not requires
independent variables to be linearly related, nor does it require equal variances within
each groups, making it a less stringent statistical analysis procedure. As a results,
logistic regression was used to predicts the likelihood of fraudulent credit cards [34].
Clarify the working of LR through the following scenario, The default variables for
determining whether a tumor is malignant or not is y = 1 (tumor malignant) the x variable

could be a measurement of the tumors, such as its size. The logistic functions convert
the x-values of the dataset's various instance’s into a range of 0 to 1. The tumor is
classified as malignant if the probability exceeds 0.5. (As indicated by the horizontal
line). As shown in the figure below:



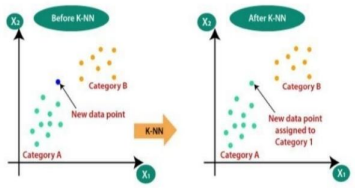
4.4 K -Nearest Neighbor

A simple, easy-to-implements supervised machine-learning technique that uses
categorized input data to develop function that gives a suitable outputs when given
additional unlabelled data. Both classifications and regression problem’s can be solved
with the k-nearest neighbors (KNN) algorithms, which is quick and straight forward to
apply. Uses labeled data to teach a functions that generates an acceptable erformances
or new data. In the K-Nearest Neighbor algorithms, the resemblance between the new
cases and the cases that are already categorized is calculated. Once the new case is
placed in a categorys that is most comparables to the available ones it is applied to all
remaining case in that groups. In an analogous fashion, KNN organizes all accessibles
data and categorizes new points depending on how similar they are.

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score, confusion_matrix

# Load your dataset containing credit card transactions
data = pd.read_csv('credit_card_data.csv')

# Preprocess the data (feature engineering, scaling, etc.)

# Split the data into training and testing sets
X_train, X_test, y_train, y_test =
train_test_split(data.drop('fraud_label', axis=1), data['fraud_label'],
test_size=0.2, random_state=42)

# Train a machine learning model, e.g., a Random Forest Classifier
model = RandomForestClassifier()
model.fit(X_train, y_train)

# Make predictions on the test set
y_pred = model.predict(X_test)

# Evaluate the model's performance
accuracy = accuracy_score(y_test, y_pred)
confusion = confusion_matrix(y_test, y_pred)
```

This describes
anytime new data emerge it is just a matter of fitting a K-N classification scheme to it.
The algorithm is very straightforward and uncomplicated to put into practice. If a model
does not needs to be built so some parameters and expectations may be tuned, it isunnecessary. The algorithms get’s significantly slower as predictors/independent
variables increase [36]. As shown in the figure below:

General structure working of the K-NN
Decide on the numbers of neighbors in the first phase (K). Definethe Euclidean distance
amidst K neighbors then locates K closest neighbors using the measured Euclidean
distance. Count the numbers of data points in every group between this KN in a
subsequent phases, then assign the modern data points to the collection with the most
neighbors. Finally, our paradigm is finished. Consider the following scenario: We
have an image of two animals: a cat and a dog, and we want to identify which one the
picture represents. As a result, the KNN can be utilized as a methods for the definition
because it is based on a likeness measures. Our KNN will look for similarities
between the latests data set and the photos of animals, andclassify it based on most
analogous attributes.



4.5 K-means Clustering
Because of its simplicity and effectiveness, it is the most widely used unsupervised
learning methodology. By calculating the mean distances between data points, this
method allocates points to groups. It then repeats this process in order to improves the
accuracy of its categorizes over time [37]. The K-Means in the figure below are

Explained via the following steps. To determine the number of clusters, choose
K. Then choose K location or centroid at random. (It could be something different from
the incoming datasets.) In the following step Assign each data points to the centroid that
is closests to it, forming the preset K clusters. Then calculates the variance and
reposition each cluster's centroid. Repeat the third step, reassigning each points to the
cluster's modern nearest centroid. Steps to finish, If there is a reassignment, go
to step 4, otherwise, move to FINISH. The model is finished.To explain how K-MC works. Consider the following situation: If hospital wishe to
establishs Care Wards. K-means Clustering will divides these high-risk location into
cluster’s and establish a cluster centre for each cluster which will be the location of the
Emergency Units. These cluster centres are each clusters centroid are located at a
minimum distance from all of the clusters points, as a results the Emergency Unit will be
located at a minimum distances from all accident-prone places within a cluster.

CONCLUSION

Credit card fraud is most common problem resulting in loss of lot money for people and loss for some banks and credit card company. This project want to help the peoples from their wealth loss and also for the banked company and trying to develop the model which more eciently separate the fraud and fraud less transaction by using the time and amount feature in data set given in the Kegel. rst we build the model using some machine learning

algorithms such as logistic regression, decision tree, support vector machine, this all are supervised machine learning algorithm in machine learning.

In feature solving this problem statement using another part of articial intelligence that is time series analysis, in our present project we used both and time and amount feature mainly for predicting the weather the transaction is fraud or Nonfraud transaction, in time series analysis we can reduce the number of parameters that is feature required for the model and we can achieve this model by using average method ,moving average or window method.